Natural Language Engineering **1** (1): 1–40. Printed in the United Kingdom

On the Interpretation of Noun Compounds: Syntax, Semantics, Entailment

PRESLAV NAKOV

Qatar Computing Research Institute, Qatar Foundation
Tornado Tower, floor 10, P.O. Box 5825
Doha, Qatar

(*Received 30 June 2012*)

Abstract

We discuss the problem of interpreting noun compounds such as *colon cancer tumor suppressor protein*, which pose major challenges for the automatic interpretation of English written text. We present an overview of the more general process of compounding and of noun compounds in particular, as well as of their syntax and semantics from both theoretical and computational linguistics viewpoint with an emphasis on the latter. Our main focus is on computational approaches to the syntax and semantics of noun compounds: we describe the problems, present the challenges, and discuss the most important lines of research. We also show how understanding noun compound syntax and semantics could help solve textual entailment problems, which would be potentially useful for a number of NLP applications, and which we believe to be an important direction for future research.

1 Introduction

"*Recent studies identify the colon cancer tumor suppressor protein adenomatous polyposis coli (APC) as a core organizer of excitatory nicotinic synapses.*" (from `http://www.tufts.edu/sackler/tcvr/overview.htm`)

An important characteristic of scientific and technical literature is the abundance of long sequences of nouns that act as a single noun, known as *noun compounds*, e.g., *bone marrow*, *web site design*, *internet connection speed test*, etc.

While eventually mastered by domain experts, noun compounds and their interpretation pose major challenges for automated analysis. For example, what is the syntactic structure of the compound *colon cancer tumor suppressor protein*: is it [[*colon cancer*] [[*tumor suppressor*] *protein*]] or [[[*colon cancer*] [*tumor suppressor*]] *protein*] or [[[[*colon cancer*] *tumor*] *suppressor*] *protein*], etc.? Can *colon cancer* be paraphrased as *cancer that occurs in the colon*? Or as *cancer in the colon*? What is the relationship between *colon cancer* and *tumor suppressor protein*? Between *colon* and *cancer*? Is a *tumor suppressor protein* a kind/type of *tumor suppressor*? Is it a kind/type of *suppressor*?

Noun compounds (NCs) cannot just be ignored by natural language processing (NLP) applications since they are abundant in English written text. Baldwin and Tanaka (2004) found that 2-4% of the tokens in various corpora are part of noun compounds: 2.6% in the *British National Corpus*, 3.9% in the *Reuters corpus*, and 2.9% in the *Mainichi Shimbun Corpus*. For example, there are 256K distinct noun compounds compared to 939K distinct wordforms in the 100M-word *British National Corpus* (BNC). Noun compounds are even more frequent in some types of text, e.g., biomedical, since they form a large part of the terminology, and in titles and abstracts, since they serve as text compression devices and allow the author to say more with fewer words. More importantly, noun compounds are very productive: NC types in English follow a Zipfian distribution (Séaghdha2008), e.g., over half of the noun compound types in BNC occur only once (Kim and Baldwin2006). A study on English new words over fifty years found that compounding is the most frequent word formation process, covering 68% of the new words; 90% of these new compounds are noun compounds (Algeo1991). This high productivity means that compounds cannot be listed in a dictionary, e.g., even for relatively frequent noun compounds occurring ten times or more in the BNC, static English dictionaries provide only 27% coverage (Tanaka and Baldwin2003). Thus, noun compounds, unless lexicalized, have to be interpreted compositionally.

Understanding noun compound syntax and semantics is difficult but potentially important for many natural language applications (NLP) including but not limited to question answering, machine translation, information retrieval, and information extraction. For example, a question answering system might need to know whether *protein acting as a tumor suppressor* can paraphrase *tumor suppressor protein*, and an information extraction system might need to decide whether *neck thrombosis* and *neck vein thrombosis* could possibly co-refer when used in the same document. Similarly, a statistical machine translation system facing the unknown noun compound *WTO Geneva headquarters* might benefit from being able to paraphrase it as *Geneva headquarters of the WTO* or as *WTO headquarters located in Geneva*. Given a query like *migraine treatment*, an information retrieval system could use paraphrasing verbs like *relieve* and *prevent* for page ranking and query refinement.

Below we discuss noun compounds, as well as the general process of compounding, from a linguistic point of view; we focus on English, but occasionally we give examples from other languages. We then discuss the syntax and semantics of noun compounds, and we show how understanding them could help textual entailment.

2 Noun Compounds in Theoretical Linguistics

In this section, we first describe the process of compounding in general – as a mechanism for producing a new word by putting two or more existing words together. Then, we focus on the special case of *noun compounds*.

This section is not meant to be an exhaustive overview of the vast volume of linguistic literature on compounding and noun compounds; we merely touch some of the most important definitions and properties of compounds in order to provide the reader with enough background before the computational discussion.

Those who would like to learn more about compounds from a theoretical linguistics viewpoint could start with an excellent short overview by Bauer (2006); if more specific details are needed, there are a number of relevant longer readings, e.g., (Downing1977; Levi1978; Warren1978; Bauer1983; Di Sciullo and Williams1987; Liberman and Sproat1992; Booij2005; Lieber and Stekauer2009). The online *Compound Noun Bibliography*[1] is another excellent reference to consider; note, however, that it mixes theoretical and computational literature.

2.1 The Process of Compounding

The *Dictionary of Grammatical Terms in Linguistics* defines the process of *compounding* as follows (Trask1993):

> "The process of forming a word by combining two or more existing words: *newspaper*, *paper-thin*, *babysit*, *video game*."

Since the process of compounding constructs new words, these words can in turn combine with other words to form longer compounds, and this process can be repeated indefinitely, e.g., *orange juice*, *orange juice company*, *orange juice company homepage*, *orange juice company homepage logo*, *orange juice company homepage logo update*, etc. Compounds of length longer than two are less frequent in general, and are more typical for technical and scientific texts.

In English, the process of compounding can combine together words belonging to various parts of speech, e.g., adjective+adjective (e.g., *dark-green*, *light-blue*), adjective+adverb (e.g., *leftmost*), adjective+noun (e.g., *hot dog*, *shortlist*, *white collar*, *highlife*), adjective+preposition (e.g., *forthwith*), adjective+verb (e.g., *highlight*, *broadcast*, *quick-freeze*, *dry-clean*), noun+adjective (e.g., *trigger-happy*, *army strong*, *bulletproof*, *dog tired*, *English-specific*, *brand-new*), noun+noun (e.g., *silkworm*, *honey bee*, *bee honey*, *stem cell*), noun+preposition (e.g., *love-in*, *timeout*, *breakup*), noun+verb (e.g., *finger-point*, *taperecord*), preposition+adjective (e.g., *overeager*, *over-ripe*), preposition+noun (e.g., *underwater*, *indoor*), preposition+preposition (e.g., *within*, *without*, *into*, *onto*), preposition+verb (e.g., *overestimate*, *withdraw*, *upgrade*, *withhold*), verb+adverb (e.g., *tumbledown*), verb+noun (e.g., *pickpocket*, *cutthroat*, *know-nothing*), verb+preposition (e.g., *countdown*, *stand-by*, *cut-off*, *castaway*), and verb+verb (e.g., *freeze-dry*). More complex structures are also possible, e.g., *state-of-the-art*, *part-of-speech* or *over-the-counter eye drop*. Note that not all of these compound types are productive in modern English.

The part of speech of an English compound is typically that of its last word, e.g., *hot dog* is a noun since its second word is the noun *dog*, while *broadcast* is a verb because it ends with the verb *cast*.[2] Some classes of compounds do not follow this rule though, e.g., noun+preposition and verb+preposition yield complex nouns instead of complex prepositions, e.g., *timeout* and *countdown*.

[1] `http://www.cl.cam.ac.uk/~do242/bibsonomy.p.html`

[2] The compound *broadcast* can be also seen as a noun, but this noun form is arguably better analyzed as derived from the verb *broadcast* through conversion.

The most typical compounds in English are those composed of nouns only, known as *noun compounds*, and the most frequent among them are those of type noun+noun (because they are the shortest). Yet, not every sequence of nouns is considered to be a noun compound.

There is little agreement in the research community on how to define the notion of *noun compound*.[3] Different authors use different definitions, focusing on different aspects, and often use different terms in order to emphasize particular distinctions. Lauer (1995) provides the following list of closely related notions used in the literature: *compound nominal*, *nominal compound*, *compound noun*, *complex nominal*, *nominalization*, *noun sequence*, *compound*, *noun compound*, *noun-noun compound*, *noun+noun compound*, *noun premodifier*. While some of these terms are broader and some are narrower in scope, most of them refer to objects that are syntactically analyzable as nouns (Chomsky and Halle1968; Jackendoff1975).

One popular definition is that of Downing (1977), who defines the notion of a *nominal compound* as a sequence of nouns which function as a single noun, e.g., *orange juice*, *company tax policy*, *law enforcement officer*, *colon cancer tumor suppressor protein*, etc. This definition is both simple and relatively unambiguous, which makes it the preferred choice for linguists working on syntax as well as for computational linguists. The primary difficulty with using it as a decision criterion for extracting noun-noun compounds in the absence of a grammar is the arbitrariness of the adjective/noun distinction for some premodifiers, e.g., are *adult* and *male* nouns or adjectives in *adult male rat*? Or, if *dental decay* and *tooth decay* are synonymous, should we then think of *tooth* as an adjective in that context? See also (Huddleston and Pullum2002) pp. 556 for a relevant discussion.

In view of these issues, some researchers have proposed a broader definition that threats some adjectival and noun modifiers alike. For example, Levi (1978) advocates the concept of *complex nominals*, which groups three partially overlapping classes: *nominal compounds* (e.g., *doghouse*, *deficiency disease*, *apple cake*), *nominalizations* (e.g., *American attack*, *presidential refusal*, *dream analysis*), and *nonpredicate NPs*[4] (e.g., *electric shock*, *electrical engineering*, *musical critism*). She argues that these subclasses share important syntactic and semantic properties. For example, the nonpredicate NP *linguistic difficulties* is arguably synonymous with the nominal compound *language difficulties*.

Many alternative definitions exist, based on various criteria, including orthographic, phonological, morphological, syntactic and semantic, none of which can clearly distinguish between noun compounds and other nominal phrases, but each pointing to some important properties shared by most compounds. Below we will explore some of these criteria, focusing on noun compounds; still, we will occasionally discuss compounds whose elements belong to other parts of speech.

[3] Under most definitions, the term *noun compound* is an example of a noun compound.

[4] Nonpredicate NPs contain a modifying adjective which cannot be used in predicate position with the same meaning. For example, the NP *solar generator* cannot be paraphrased as **generator that is solar*.

2.1.1 Orthographic Criteria

One simple criterion is based on orthography: many compounds are at least partially lexicalized and as such are written as a single word or hyphenated, e.g., *silkworm*, *snowball*, *steamboat*, *healthcare*, *cut-off*, *stand-by*, *coach-player*.

While concatenated orthography is a very reliable *indicator* (but not criterion!) of compoundness, hyphenated spelling is less so, e.g., *US-China* is not a compound. In contrast, *US-China relations* is a compound despite involving a separate word.

Overall, orthography is not a good criterion for English since concatenation is rare, except for short and established lexicalizations such as *textbook*, *newspaper*, *homework*, *Sunday*, while hyphenated forms are mainly used for some special kinds of compounds (defined below), such as copulative (e.g., *Bosnia-Herzegovina*) and appositional (e.g., *coach-player*, *member-state*), as well as between two of the nouns (typically the first two nouns) in some noun compounds of length three or longer, e.g., *law-enforcement officer*.

Another reason for this criterion not being very reliable for English is variation in spelling: often the same compound can appear orthographically separated, connected with a hyphen, or concatenated, e.g., *health care*, *health-care*, *healthcare*. It would be inconsistent to believe that *healthcare* and *health-care* are noun compounds, while *health care* is not.

While the orthographic criterion is of some utility for English, even if limited, it is not applicable at all to languages where the writing system does not indicate the word boundaries, as in Chinese, Japanese and Vietnamese, or for languages that have no writing system.

However, the criterion is quite useful for Germanic languages, where noun compounds are almost exclusively concatenated,[5] e.g., *nagellackborttagningsmedel* (Swedish, "nail polish remover"), *sannsynlighetsmaksimeringsestimator* (Norwegian, "maximum likelihood estimator"), *kvindehåndboldlandsholdet* (Danish, "the female handball national team"), *Sprachgruppe* (German, "language group"), *wapenstilstandsonderhandeling* (Dutch, "ceasefire negotiation").

In these languages, orthography is quite consistent with compoundness, but problems could still arise, e.g., because some Dutch speakers tend to write compounds separately under English influence. Typos are another problem and can make a big difference in meaning, e.g., in Norwegian, the concatenated form *lammekoteletter* is a compound meaning "lamb chops", while *lamme koteletter* is not a compound and means "lame, or paralyzed, chops"; similarly, the compound *røykfritt* means "no smoking", lit. "smoke free", while the noncompound *røyk fritt* has the opposite meaning of "smoke freely".

[5] Due to the concatenated spelling, compounds can get very long, e.g., the longest German word in actual use is *Rindfleischetikettierungsüberwachungsaufgabenübertragungsgesetz*, meaning "beef labelling supervision duty assignment law". It is composed of the following nouns: *Rind* (cattle), *Fleisch* (meat), *Etikettierung(s)* (labelling), *Überwachung(s)* (supervision), *Aufgaben* (duties), *Übertragung(s*) (assignment) and *Gesetz* (law).

2.1.2 Phonological Criteria

Chomsky and Halle (1968) gave a phonological definition for noun compounds in English: the words preceding a noun will form a compound with it if they receive the primary stress. Therefore, *blackboard* is a compound (*black* gets the primary stress), but *black board* is not (equal stress).

Knowing whether a sequence of words represents a compound might be useful for a speech synthesis system, where using the wrong stress can convey a meaning that is different from what is intended. For example, using fronted stress would make *French teacher* a compound with the meaning of *teacher who teaches French*, while double stress would convey the noncompound meaning of *teacher who is French* (Levi1978). For compounds longer than two words, the correct pronunciation also depends on their internal syntactic structure, which makes a noun compound parser an indispensable component of an ideal speech synthesis system.

If the internal syntactic structure is known, the stress assignment algorithm of Chomsky et al. (1956) can be used, which follows the constituent structure from the inside out, making the stress of the first constituent primary and reducing the rest by one degree. For example, if *black board* is not a compound, each word would get primary stress [*black/1*] [*board/1*], but if it is a compound then the stress of the second word would be reduced: [*black/1 board/2*]. Now, if [*black/1 board/2*] is a sub-constituent of a longer compound, e.g., *black board eraser*, the rule would be applied one more time, yielding [[*black/1 board/3*] *eraser/2*], etc.

Chomsky and Halle's (1968) definition is appealing since it stems from a standard linguistic test of whether a sequence of lexemes represents a single word, and is consistent with the definition of Trask (1993) from Section 2.1, which views the process of compounding as a new word formation process.

However, it is also problematic for a number of reasons. First, stress could differ across dialects and even across speakers of the same dialect. It could also separate semantically parallel examples, e.g., according to stress patterns, *apple cake* would be a noun compound, while *apple pie* would not be. Moreover, the stress criterion is of limited use for most computational approaches to language analysis: they work primarily with written text, where no information about the stress is available.

It has been also pointed out that there are other factors involved. For example, *Oxford Street*, has fronted stress, i.e., on *Oxford*, while *Oxford Road* gets stressed on *Road*. This is arguably due to the fact that *Street* is a common and expected term in the category of street names, and as such it carries little semantic information, while *Road* is a more specific and less expected term and is to be stressed (Ladd1984).

More generally, it has been proposed that the fronted stress in compounds is merely a result of deaccenting the head noun, where the stress would go normally in noun phrases (Ladd1984). The head is deaccented when the modifier subcategorizes the head, in which case the head provides only part of what is needed to identify the semantic category of the whole. For example, *green house* with stress on *house* means *house that is green*, which is merely a specification of the more general category *house*, while fronted stress yields a new category, and thus here the modifier *green* provides crucial new information, which justifies it being stressed.

However, nonfronted stress is also frequent for compounds. Liberman and Sproat (1992) find that 25% of the two-word noun compounds in their corpus have a main stress on the right. Most of them fall under particular semantic relations such as: (a) things made of/with something (as a part), e.g., *steel plate*, *chicken soup*, (b) place where something is found, e.g., *garage door*, *mountain pass*, *city employee*, (c) time when something is found, e.g., *summer palace*, *fall weather*, *spring flowers*, (d) proper name modifiers, e.g., *US ambassador*, *Carlsberg beer*, *Napoleon brandy*, (e) left-headed compounds, e.g., *vitamin D*, *peach melba*, *planet Earth*, and some other, e.g., *barrier reef*, *child labor*, *World Bank*.

Stress criteria can be applied to some other languages as well. For example, in Norwegian, the front-stressed form *smørbrød* means "sandwich", while the normal stress form *smør brød* means "butter bread" as a command.

2.1.3 Morphological Criteria

Another criterion is based on the idea that noun compounds are single lexemes and should behave as such morphologically. For English, this means that a compound should inflect as a whole only, while compound-internal inflections should not be allowed. For example, *apple cake* can inflect for plural to become *apple cakes*, but it cannot inflect internally, i.e., **apples cake* or **apples cakes* are not possible. Moreover, being a single lexeme, the compound is often free to inflect in a way that is different from the way its last word would, e.g., *sabre tooth* (a type of pre-historic tiger) forms a plural as *sabre tooths* and not as **sabre teeth*.

There are counter-examples though, e.g., compounds like *suggestions poll*, *weapons treaty*, *students commission* are readily composed with internal plural.

Counter-examples can be found in other languages as well. For example, in Bulgarian the definite article is a clitic that attaches to the first suitable word in a noun phrase (e.g., a noun, an adjective, a participle, etc.); this also applies to compounds such as *kyshta-muzey* ("house museum"), which inflects for definiteness as *kyshtata-muzey* (lit. "house-the museum", meaning "the house museum").

In Russian, the inflection can be unstable, and could target one of the nouns only, as well as both nouns simultaneously, e.g., the prepositional case inflection for *vagon-restoran*, (lit. "car-restaurant", meaning "dining car in a train") can be *v vagone-restorane* (preferred in writing), or *v vagon-restorane* (preferred in spoken language); it is not uncommon to find also *v vagone-restoran*.

There are some languages, however, for which the morphological criterion works very well. One example is Turkish, which has three types of noun compounds, two of which are marked morphologically by a special suffix. Type 1 shows ownership, e.g., *Ayşe'nin kitabı* (lit. "Ayshe-gen book-poss", meaning "Ayshe's book"), where *Ayşe* is marked with a genitive case, and *kitap* takes a possessed suffix. As in English,[6] the modifier does not have to be a person's name, e.g., *manavın merakı* is formed out of *manav* ("greengrocer") + *merak* ("curiosity").

[6] Note that English genitive constructions are not considered compounds.

Type 2 includes noun compounds that do not involve possession: thus, there is no genitive, but the possessed suffix stays, e.g., *göbek dansı* ("belly dance") is formed out of *göbek* + *dans*. Type 3 includes noun compounds, where both nouns are in nominative, i.e., not inflected; such compounds can express what the second element is made of, e.g., *cam bardak* ("glass tumbler", lit. "tumbler glass"), or what it resembles, e.g., *canavar adam* ("man-beast", lit. "beast man").

Other languages can sometimes, but not always, signal compounding morphologically with a special linking element, e.g., *parohod* in Russian (lit. "steam boat", meaning "steamboat"), *Tagebuch* in German (lit. "day book", meaning "diary"), *chemin-de-fer* in French (lit. "way of iron", meaning "railway").

2.1.4 Syntactic Criteria

Syntactic criteria check whether a compound is treated as a single unit in syntax, i.e, whether syntax can "see" the individual words that form it. Bauer (2006) suggests that one way to test this is by trying to construct a sentence where a pronoun would target the individual words in the compound instead of the compound as a whole. For example, we can say *It's a plastic bag, not a paper one.*, where *one* refers to *bag*, not to *plastic bag*, which shows that *plastic bag* is transparent to syntax, and thus it should not be considered a compound, according to this criterion. On the other hand, we cannot produce similar kinds of sentences for compounds like *headache*, e.g., a sentence like **I have a headache, not a stomach one.* is not possible, which means that *headache* should be considered a noun compound. This is not a bullet-proof criterion as people do sometimes generate "impossible" sentences.

This criterion is related to the morphological criterion in Section 2.1.3, e.g., the definiteness marker in the Bulgarian compound *kyshtata-muzey*, which involves an inflectional clitic that targets an individual word, demonstrates that this compound is transparent to syntax, since clitics are syntactic objects.

2.1.5 Semantic Criteria

Various semantic criteria for compoundness have been proposed in the literature; here we will consider *permanence*, *noncompositionality* and *lexicalization*.

The permanence criterion requires that the words forming a compound be in a permanent or at least habitual relationship, e.g., *desert rat* can only refer to a rat that is strongly associated with a desert, e.g., living in or around it. Unfortunately, this criterion rejects many arguably good noun compounds such as *heart attack* and *birth trauma*, which describe short episodic events, not permanent ones.

The noncompositionality criterion asks that compounds be at least partially non-compositional. Bauer (2006) gives the examples of *wheel-chair* and *push-chair*: while both can have wheels and can be pushed, each of them has a particular specialization of meaning. Compositionlity is a matter of degree, which makes it problematic as a criterion. Some noun-noun compounds are completely noncompositional, e.g., *honeymoon* has nothing to do with *honey* or *moon*. Other examples could be argued to be productively composed, e.g., *orange peel*.

Many other compounds lie in the continuum in between, e.g., *boy friend* and *healthcare* exhibit a lower degree of compositionality. Towards the other end are the metaphorical *ladyfinger* and *birdbrain*, which are highly compositional.

The last criterion, lexicalization, asks that noun compounds be at least partially lexicalized, i.e., that they tend to represent single lexical entries. This criterion is highly correlated with noncompositionality as noncompositional compounds tend to be also lexicalized, which is often signalled by the variation in spelling. For example, both *health care* and *healthcare* are commonly used, but the latter suggests a higher degree of lexicalization compared to the space-separated version. Similarly, the concatenated *bathroom* is more lexicalized than *game room*. In some cases, a high degree of lexicalization can be signalled by spelling changes in the compounded form, e.g., *dough* + *nut* = *donut*.

2.2 Types of Compounds

The two most important types of noun compounds are *endocentric* and *exocentric*. The *Lexicon of Linguistics*[7] defines them as follows:

Endocentric compound: a type of compound in which one member functions as the head and the other as its modifier, attributing a property to the head. The relation between the members of an endocentric compound can be schematized as "*AB* is (a) *B*". Example: the English compound *steamboat* as compared with *boat* is a modified, expanded version of boat with its range of usage restricted, so that *steamboat* will be found in basically the same semantic contexts as the noun *boat*. The compound also retains the primary syntactic features of *boat*, since both are nouns. Hence, a *steamboat* is a particular type of *boat*, where the class of *steamboats* is a subclass of the class of *boats*.

Exocentric compound: a term used to refer to a particular type of compound, viz. compounds that lack a head. Often these compounds refer to pejorative properties of human beings. A Dutch compound such as *wijsneus* "wise guy" (lit. "wise-nose") (in normal usage) does not refer to a nose that is wise. In fact, it does not even refer to a nose, but to a human being with a particular property. An alternative term used for compounds such as *wijsneus* is *bahuvrihi* compound.

There is a third category of compounds, known as *copulative* or *coordinative* (*dvandva* in Sanskrit). They form an entity that is the sum of the two nouns that form the compound and is also distinct from either of them. Copulative compounds are rare in English; examples include mostly hyphenated proper nouns, e.g., *Austria-Hungary* and *Bosnia-Herzegovina*, but also some common nouns, e.g., *gerund-participle*. However, they are very common in Indic languages (e.g., Sanskrit, Hindi, Urdu, Marathi, and Tamil), as well as in Chinese and Japanese.

Copulative compounds are often confused with a fourth category of compounds known as *appositional*, where each of the nouns denotes independently a different aspect of the entity that the compound represents, e.g., *coach-player*, *sofa-bed*, *writer-director*, *learner-driver*, and *programmer analyst*. For example, *coach-player* is somebody who is both a coach and a player.

[7] `http://www2.let.uu.nl/Uil-OTS/Lexicon/`

Noun compounds can be also formed by reduplication. There are various kinds of reduplication in English such as exact (e.g., *bye-bye*), ablaut (e.g., *chit-chat*) rhyming (e.g., *walkie-talkie*, *hokey-pokey*), sh-reduplication (e.g., *baby-shmaby*), contrastive (e.g., *I'll make the tuna salad, and you make the salad-salad.*), etc. Despite this variety, reduplication is not very common nor it is very productive in English, except probably for the last two categories. However, it is very common in some other languages, e.g., in Malay it is used, among many other things, for forming plural, e.g., *ayam-ayam*[8] "chickens" is the plural of *ayam* "chicken".

Finally, there are *portmanteaux* compounds, which are composed by blending the sounds of two or more words, while also combining their meanings. For example, *brunch* is formed from *breakfast + lunch*, while *Eurasia* blends together *Europe* and *Asia*. A more recent example is *Merkozy*, which blends *Merkel* and *Sarkozy*. Portmanteaux are frequent in Russian, where they used to be the preferred choice for Soviet-style communist and propaganda terms, e.g., *komsomol* is made of *kom*munisticheskiy *so*yuz *mo*lodyozhi, meaning "Young Communist League", or *sohkhoz*, from *sov*etskoe *khoz*yaistvo, meaning "Soviet farm".

2.3 Properties of Compounds

2.3.1 Headedness

An important characteristic of compounds is their headedness.

For example, in *endocentric compounds* such as *birdcage*, one member functions as the head and the other as its modifier, attributing a property to the head. In contrast, exocentric compounds such as *birdbrain* lack an overtly expressed semantic head, and are thus headless in semantic sense. Finally, copulative compounds such as *Bosnia-Herzegovina* and *gerund-participle*, as well as appositional compounds such as *coach-player* and *sofa-bed* have two semantic heads.

Endocentric noun compounds are predominantly right-headed in English, e.g., the head of *birdcage* is *cage*, which is the right noun: a *birdcage* is a kind of *cage*. However, some English compounds are left-headed, e.g., *vitamin D*, which is a kind of *vitamin*, not a kind of *D*. Other examples include constructions where the second word is an identifying name or a number, e.g., *Cafe Vienna*, *interferon alpha*, *exit 15*, *Route 66*, *Linguistics 101*, borrowings from languages like French, which is mostly left-headed, e.g., *beef Julienne*, and compounds where the first noun is a classifier, e.g., *Mount Whitney*, *planet Earth*, *President Obama*.

Thus, the order of the words that form a noun compound is very important for its semantics. Take for example *birdcage*, which is a kind of cage: if we switch the order of *cage* and *bird*, we get *cagebird*, which is a valid noun compound, but its meaning is quite different – it has *bird* as a head, i.e., a *cagebird* is a kind of *bird*. In some cases, switching the order might not even be possible, e.g., *healthcare* is a valid compound, but *care health* is harder to interpret.

[8] Note that *ayam-ayam* is not a compound; it is a plural form formed by reduplication.

Other languages can have a different pre-dominant order of the modifier and the head in endocentric noun compounds, which order typically follows the head-modifier order in the noun phrases of that language. For example, noun phrases in Romance languages are mostly left-headed, e.g., *urânio enriquecido* (Portuguese, lit. "uranium enriched", i.e., "enriched uranium"). The same principle naturally extends from adjective-noun modification to noun compounds, e.g., *estado miembro* (Spanish, lit. "state member", meaning "member-state"), or *legge quadro* (Italian, lit. "law framework", meaning "framework law").

2.3.2 Transparency

We have already discussed lexicalization as a semantic criterion in Section 2.1.5. A related property of compounds is transparency. For example, even though highly lexicalized, *ladyfinger* can be analyzed as "a pastry that resembles a lady finger".

Levi (1978) arranges compounds on a transparency scale: (1) transparent, e.g., *mountain village*, *orange peel*, (2) partly opaque, e.g., *grammar school*, *brief case*, (3) exocentric, e.g., *birdbrain*, *ladybird*, (4) partly idiomatic, e.g., *monkey wrench*, *flea market*, and (5) completely idiomatic, e.g., *honeymoon*, *duck soup*.

Lexicalization and transparency are related, but different notions: nontransparent compounds such as *honeymoon* are necessarily lexicalized, but the converse is not true. For example, *Sunday* is highly lexicalized but partially transparent: one can easily see "*sun* + *day*". *Sunday* is not fully transparent though since for full transparency the words have to be used in their usual way, and it is not evident to a modern speaker of English that there is any connection between *Sunday* and the *sun*, other than the word itself.

Another class of opaque compounds includes those borrowed from a foreign language as readily constructed, e.g., the English *hippopotamus* comes from the Greek *hippopotamos*, which is an alteration of *hippos potamos*, i.e., "*river horse*".

2.3.3 Syntactic Ambiguity

Noun compounds of length three or longer can have multiple interpretations due to structural ambiguity. For example, *plastic water bottle* is ambiguous between a left- and a right-bracketing:

(1) [[*plastic water*] *bottle*]
(2) [*plastic* [*water bottle*]]

The correct interpretation is (2), meaning *water bottle that is made of plastic*, while in (1) we have a *bottle* that has something to do with "*plastic water*".

In spoken English, the correct interpretation will be signalled by the stress pattern used by the speaker, as we have discussed in Section 2.1.2. Another possible way to signal the structure at speech time would be to add a pause at the appropriate position: *plastic water* _ _ *bottle* vs. *plastic* _ *water bottle*.

Some languages have built-in mechanisms for eliminating most instances of such kinds of structural ambiguities. Let us take Turkish noun compounds of Type 2 (see Section 2.1.3) as an example. A noun compound like *göbek dansı* functions as a single noun and can be used as a modifier of another noun, e.g., *kurs* ("course"), which yields the left-bracketed double-marked noun compound *göbek dansı kursu* ("belly dance course"). In contrast, a right-bracketed noun compound like *Ankara mangal partisi* ("Ankara BBQ party") has only one possessed suffix. Its derivation first puts together *mangal* and *parti* to form the noun compound *mangal partisi*, which in turn is modified by *Ankara*. Note that *parti* does not acquire a second suffix in the process since Turkish does not allow double possessed suffixes on the same noun; in some cases, this can cause ambiguity for longer compounds, e.g., *Ankara göbek dansı kursu* is ambiguous between the following two bracketings:

(3) [[*Ankara* [*göbek dansı*]] *kursu*]
(4) [*Ankara* [[*göbek dansı*] *kursu*]]

2.3.4 Language-dependency

Noun compounds are inherently language-dependent: what is considered a compound in one language might not be a compound in another language for language-internal reasons. For example, we have seen that structures involving genitive modifiers such as *Ayşe'nin kitabı* are compounds in Turkish, but not in English, e.g., *Ayshe's book* is not a compound. This is because the genitive case in English is formed using a clitic rather than a suffix.

In many languages, the relationship between the modifier and the head is expressed using a variety of grammatical cases. For example, the Sanskrit *tatpurusa* compounds involved cases such as genitive (e.g., *raja-putra*, lit. "king-son", meaning "son of a king"), accusative (e.g., *jaya-prepsu*, "victory-desiring"), dative (e.g., *visnu-bali*, lit. "Vishnu-offering", meaning "offering to Vishnu"), instrumental (e.g., *deva-datta*, lit. "god-given", meaning "given by the gods"), locative (e.g., *purvahna-krta*, lit. "morning-done", meaning "done in the morning"), and ablative (e.g., *svarga-patita*, lit. "heaven-fallen", meaning "fallen from heaven").

Head-modifier structures parallel to those for Sanskrit are not considered compounds in Russian; they are called *word combinations*. As in Sanskrit, they can use a variety of cases such as genitive (e.g., *kniga uchenika*, lit. "book student-gen", i.e., "student's book"), dative (e.g., *pamyatnik partizanam*, lit. "monument partisans-dat", i.e., "monument to the partisans"), and instrumental (e.g., *porez nozhom*, lit. "cut knife-instr", i.e., "cut with a knife").

3 Noun Compound Syntax

Here we switch from theoretical discussion on the definition and properties of noun compounds to computational work on noun compound interpretation.

The semantic interpretation of noun compounds of length three or more requires that their syntactic structure be determined first. Consider for example the following contrastive pair of noun compounds:

(1) *liver cell antibody*
(2) *liver cell line*

In example (1), there is an *antibody* that targets a *liver cell*, while example (2) refers to a *cell line* that is derived from the *liver*. In order to make these semantic distinctions accurately, it can be useful to begin with the correct grouping of terms, since choosing a particular syntactic structure limits the options left for semantics. Although parallel at the part-of-speech (POS) level, the above two noun compounds have different constituency trees, as Figure 1 shows.

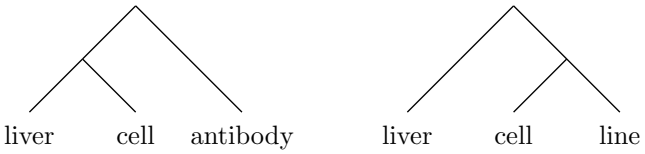

Fig. 1. **Left vs. right bracketing:** constituency trees.

The trees on Figure 1 can be represented using brackets, which gives the name of the task, *noun compound bracketing* (we assume right-headedness in this section):

(1b) [[*liver cell*] *antibody*] (left bracketing)
(2b) [*liver* [*cell line*]] (right bracketing)

Longer noun compounds like *colon cancer tumor suppressor protein* are rarely dealt with, probably because they are relatively rare in any case, with the exception of scientific texts. Instead, parsing them has typically been reduced to similar left/right-bracketing decisions for triples of nouns. For example, suppose we have decided that [*colon cancer*] and [*tumor suppressor*] are noun compounds used as subunits in the bracketing: [[*colon cancer*] [*tumor suppressor*] *protein*]. Assuming a noun compound behaves like its head, we end up with a bracketing problem for the compound *cancer suppressor protein*. If we decide on a right bracketing for that compound, we end up with the following overall structure:

[[*colon cancer*] [[*tumor suppressor*] *protein*]]

Noun compound parsing is a necessary step when performing semantic interpretation since the syntactic structure reveals the sub-parts between which relations need to be assigned, e.g., for the above example, we can have the following semantic representation:

[*tumor suppressor protein*] which is implicated in [*colon cancer*] (`IN`; `LOCATION`)
[*protein*] that acts as [*tumor suppressor*] (`IS`; `AGENT`)
[*suppressor*] that inhibits [*tumor(s)*] (`OF`; `PURPOSE`)
[*cancer*] that occurs in [*(the) colon*] (`OF`; `IN`; `LOCATION`)

The best known early work on automated unsupervised noun-compound bracketing is that of Lauer (1995), who introduced the probabilistic *dependency model* and criticized the previously used *adjacency model* (Marcus1980; Pustejovsky et al.1993; Resnik1993). Given a three-word noun compound $w_1w_2w_3$, the adjacency model compares the strengths of association $Assoc(w_1, w_2)$ and $Assoc(\underline{w_2}, w_3)$, while the dependency model compares $Assoc(w_1, w_2)$ and $Assoc(\underline{w_1}, w_3)$.

3.1 Adjacency Model

According to the bracketing representation introduced above, given a three-word noun compound $w_1w_2w_3$, the task is to decide whether w_2 is more closely associated with w_1 or with w_3. Therefore, it looks natural to compare the strength of association between the first two and the last two words, which yields an **adjacency model** (Lauer1995):

- if $Assoc(w_1, w_2) < Assoc(\underline{w_2}, w_3)$, predict *right bracketing*;
- if $Assoc(w_1, w_2) = Assoc(\underline{w_2}, w_3)$, make no prediction;
- if $Assoc(w_1, w_2) > Assoc(\underline{w_2}, w_3)$, predict *left bracketing*.

3.2 Dependency Model

Lauer (1995) proposed an alternative, syntactically motivated *dependency model*. It is illustrated in Figure 2, where arcs point from the heads to their modifiers.[9]

Fig. 2. **Left vs. right bracketing:** dependency structures.

In this representation, both the left and the right dependency structures contain a link $w_2 \leftarrow w_3$, but differ because of $w_1 \leftarrow w_2$ and $w_1 \leftarrow w_3$, respectively.[10] Therefore, the **dependency model** focuses not on w_2 but on w_1:

- if $Assoc(w_1, w_2) < Assoc(\underline{w_1}, w_3)$, predict *right bracketing*;
- if $Assoc(w_1, w_2) = Assoc(\underline{w_1}, w_3)$, make no prediction;
- if $Assoc(w_1, w_2) > Assoc(\underline{w_1}, w_3)$, predict *left bracketing*.

[9] Lauer (1995) had in his formulas arrows pointing in the opposite direction, i.e., from the modifier to the head, but here we have opted to use the standard notation for dependency parsing, where arcs point from the head to its modifier.

[10] In these examples, the arcs always point to the left, i.e., the head always follows the modifier. While this is the typical case for English, there are some rare cases where it does not hold, e.g., in *Hepatitis B* the head is *Hepatitis*; we discussed this in Section 2.3.1.

3.3 Adjacency vs. Dependency

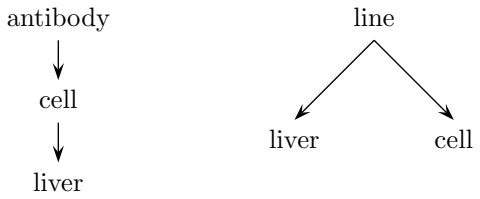

Fig. 3. **Left vs. right bracketing:** dependency trees.

Figure 3 shows the *dependency trees* corresponding to the dependency structures from Figure 2. Note the structural difference from the constituency trees in Figure 1. First, the constituency trees contain words in the leaves only, while the dependency trees have words in the internal nodes as well. Second, the constituency trees are *ordered binary trees*: each internal node has exactly two *ordered* descendants, one left and one right, while there is no such ordering for the dependency trees.

Consider also examples (3) and (4), which are both right-bracketed, but the order of the first two words is switched (we consider *adult* and *male* as nouns here).

(3) [*adult* [*male rat*]] (right bracketing)
(4) [*male* [*adult rat*]] (right bracketing)

Despite (3) and (4) being different, the corresponding dependency structures are equivalent as Figures 4 and 5 show: there are no dependency arcs between *adult* and *male*, and therefore changing their linear order does not alter the dependency structure.

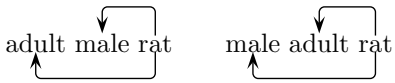

Fig. 4. Dependency structures for *adult male rat* and for *male adult rat*.

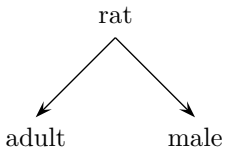

Fig. 5. Shared dependency tree for *adult male rat* and for *male adult rat*.

The adjacency and the dependency models target different kinds of right-bracketed structures. Consider for example *home health care*: *health care* is a compound, which in turn is modified by *home* as a whole, which can easily be seen in the concatenated spelling *home healthcare*, where the last two words are written without separating space.

This is different from *adult male rat*, where the order of the modifiers *adult* and *male* could be switched, which suggests that *adult* and *male* independently modify *rat*. We can conclude that, given a three-word noun compound $w_1w_2w_3$, there are two reasons it may be right-bracketed, $[w_1[w_2w_3]]$:

(a) w_2w_3 is a compound, which is modified by w_1;
(b) w_1 and w_2 independently modify w_3.

Let us now look closely at the *adjacency* and the *dependency* models:

- **adjacency model:** compare $Assoc(w_1, w_2)$ and $Assoc(\underline{w_2}, w_3)$;
- **dependency model:** compare $Assoc(w_1, w_2)$ and $Assoc(\underline{w_1}, w_3)$.

We can see that the adjacency model checks (a), i.e., whether w_2w_3 is a compound (a test for lexicalization), while the dependency model checks (b), i.e., whether w_1 modifies w_3 (a test for syntactic modification).

Note that there is only a modificational choice in case of left bracketing. If w_1 modifies w_2, then w_1w_2 is a noun compound, which now acts as a single noun to modify w_3. Consider for example *law enforcement agent*: *law* is a modifier of *enforcement*, together they form a noun compound *law enforcement*, which in turn, as a whole, modifies *agent*.

3.4 Frequency-based Association Scores

The simplest association score measures the strength of association as a bi-gram frequency in some corpus:

$$Assoc(w_i, w_j) = \#(w_i, w_j) \tag{1}$$

Another association score is based on conditional probability:

$$Assoc(w_i, w_j) = \Pr(w_i \leftarrow w_j | w_j) \tag{2}$$

$\Pr(w_i \leftarrow w_j|w_j)$ is the probability that w_i modifies the head w_j $(1 \leq i < j \leq 3)$. This probability can be estimated as follows:

$$\Pr(w_i \leftarrow w_j|w_j) = \frac{\#(w_i, w_j)}{\#(w_j)} \tag{3}$$

where $\#(w_i, w_j)$ and $\#(w_j)$ are the corresponding bigram and unigram frequencies.

Other popular association measures, based on similar unigram and bigram statistics, include *pointwise mutual information* and *Chi square*. The n-gram statistics are typically collected from a large text corpus or from the Web.

For example, Lauer (1995) collected n-gram statistics from *Grolier's encyclopedia*[11], which contained about eight million words (in 1995). To overcome data sparseness issues, he estimated the probabilities over conceptual categories t_i in *Roget's thesaurus*[12] rather than for individual words.

[11] `http://go.grolier.com`
[12] `http://thesaurus.reference.com/`

For evaluation, he created a dataset of 244 three-word noun compounds extracted from Grolier's encyclopedia; this is the benchmark testset for the task of noun compound bracketing. Most compounds in this dataset are left-bracketed: 66.80%. Lauer achieved 68.90% and 77.50% accuracy with the adjacency and the dependency models, respectively. He further reported 80.70% with a better tuned system.

More recently, Lapata&Keller (2004; 2005) estimated n-gram counts from the 100M-word British National Corpus (BNC) and from AltaVista, using page hit counts as a proxy for n-gram frequencies. They reported 68.03% accuracy with BNC and 78.68% accuracy with AltaVista.

Girju et al. (2005) used fifteen semantic features derived from WordNet (Fellbaum1998), achieving 83.10% accuracy. Unlike the above approaches, their approach is supervised and requires that each noun be annotated with the correct WordNet sense, which is an unrealistic assumption.

Nakov & Hearst (2005a) used Web-derived surface features and paraphrases,[13] achieving 89.34% accuracy on the Lauer dataset, and 95.35% accuracy on their own biomedical dataset.[14] Their most important features include the following:

- **Dashes:** e.g., finding on the Web the dashed form *law-enforcement officer* suggests a left bracketing interpretation for *law enforcement officer*.
- **Genitives:** e.g., finding the genitive form *math problem's solution* indicates a left bracketing for *math problem solution*.
- **Internal capitalization:** e.g., finding on the Web *Plasmodium vivax Malaria* suggests left bracketing for *plasmodium vivax malaria*.
- **Internal inflection variety:** e.g., *bone minerals density* favors left bracketing for *bone mineral density*.
- **Internal slashes:** e.g., *leukemia/lymphoma cell* suggests right bracketing for *leukemia lymphoma cell*.
- **Parentheses:** e.g., finding on the Web *growth factor (beta)* indicates left bracketing for *growth factor beta*.
- **Abbreviations:** e.g., finding *tumor necrosis factor (NF)* signals right bracketing for *tumor necrosis factor*.
- **Concatenations:** e.g., finding on the Web *healthcare reform* favors left bracketing for *health care reform*.
- **Word reordering:** e.g., finding *male adult rat* indicates right bracketing for *adult male rat*.
- **Paraphrases:** They can indicate the bracketing structure by showing which two words would be kept together, e.g., *cells in the bone marrow* suggests left bracketing for *bone marrow cell* since it keeps the two left words together.

[13] These features were found useful not only for bracketing but also for some other structural ambiguity problems (Nakov and Hearst2005c). See also (Nakov and Hearst2005b) for a discussion on the stability of page hits when used as a proxy for n-gram frequencies.
[14] This dataset and more details can be found in an appendix in (Nakov2007).

RDP	Example	Subj/obj	Traditional Name
$\texttt{CAUSE}_1$	*tear gas*	object	causative
$\texttt{CAUSE}_2$	*drug deaths*	subject	causative
$\texttt{HAVE}_1$	*apple cake*	object	possessive/dative
$\texttt{HAVE}_2$	*lemon peel*	subject	possessive/dative
$\texttt{MAKE}_1$	*silkworm*	object	productive/composit.
$\texttt{MAKE}_2$	*snowball*	subject	productive/composit.
USE	*steam iron*	object	instrumental
BE	*soldier ant*	object	essive/appositional
IN	*field mouse*	object	locative
FOR	*horse doctor*	object	purposive/benefactive
FROM	*olive oil*	object	source/ablative
ABOUT	*price war*	object	topic

Table 1. **Levi's recoverably deletable predicates (RDPs).** *Column 3 shows the modifier's function in the corresponding paraphrasing relative clause: when the modifier is the subject of that clause, the RDP is marked with the index 2.*

4 Noun Compound Semantics

We first present a brief overview of the representations of noun compound semantics in both theoretical and computational linguistics. Then, we describe the two main lines of research on automatic semantic interpretation of noun compounds in NLP.

4.1 Representation

Most work on noun compound semantics has focused on two-word noun compounds, or noun-noun compounds, which are the most common type. In terms of representation, in computational linguistics, the semantics of a noun compound has been typically expressed by an abstract relation like `CAUSE` (e.g., *malaria mosquito*), `SOURCE` (e.g., *olive oil*), or `PURPOSE` (e.g., *migraine drug*), drawn from a small fixed inventory. Such inventories were initially proposed by linguists, e.g., based on corpus studies, but they have become especially popular in computational linguistics.

Warren (1978) proposed a linguistic theory based on a study of the Brown corpus (Kucera and Francis1967). In this theory, noun compounds are characterized by abstract semantic relations organized into a four-level hierarchy, where the top level is occupied by six coarse-grained relations, `Possession`, `Location`, `Purpose`, `Activity-Actor`, `Resemblance`, and `Constitute`, which are further subdivided into finer-grained relations. For example, `Constitute` is divided into three level-2 relations: `Source-Result`, `Result-Source` or `Copula`. `Copula` in turn is sub-divided into another three level-3 relations `Adjective-Like_Modifier`, `Subsumptive`, and `Attributive`. Finally, `Attributive` is divided into two level-4 relations `Animate_Head` (e.g., *girl friend*) and `Inanimate_Head` (e.g., *house boat*).

	Subjective	**Objective**	**Multi-modifier**
Act	*parental refusal*	*dream analysis*	*city land acquisition*
Product	*clerical errors*	*musical critique*	*student course ratings*
Agent	—	*city planner*	—
Patient	*student inventions*	—	—

Table 2. **Levi's nominalization types with examples.**

In the alternative linguistic theory of Levi (1978), noun compounds (and complex nominals in general, see Section 2.1) can be derived by the following two processes:

1. **Predicate Deletion.** It can delete the 12 abstract recoverably deletable predicates (RDPs) shown in Table 1, e.g., *pie made of apples* → *apple pie*. In the resulting nominals, the modifier is typically the object of the predicate; when it is the subject, the predicate is marked with the index 2.
2. **Predicate Nominalization.** It produces nominals whose head is a nominalized verb, and whose modifier is derived from either the subject or the object of the underlying predicate, e.g., *the President refused General MacArthur's request* → *presidential refusal*. Multi-modifier nominalizations retaining both the subject and the object as modifiers are possible as well. Therefore, there are three types of nominalizations depending on the modifier, which are combined with the following four types of nominalizations the head can represent: *act*, *product*, *agent* and *patient*. See Table 2 for examples.

These linguistic theories have influenced the efforts to create similar inventories for computational linguistic purposes. For example, Ó Séaghdha (2007) revised the inventory of Levi (1978) based on six pragmatic criteria:

- the inventory of relations should have good coverage;
- relations should be disjunct, and should describe a coherent concept;
- the class distribution should not be overly skewed or sparse;
- the concepts underlying the relations should generalize to other linguistic phenomena;
- the guidelines should make the annotation process as simple as possible;
- the categories should provide useful semantic information.

This yielded an inventory of six semantic relations, which are further subdivided into sub-categories: `BE` (`identity`, `substance-form`, `similarity`), `HAVE` (`possession`, `condition-experiencer`, `property-object`, `part-whole`, `group-member`), `IN` (`spatially located object`, `spatially located event`, `temporarily located object`, `temporarily located event`), `ACTOR` (`participant-event`, `participant-participant`), `INST` (`participant-event`, `participant-participant`), `ABOUT` (`topic-object`, `topic-collection`, `focus-mental activity`, `commodity-charge`). E.g., *tax law* is `topic-object`, *crime investigation* is `focus-mental activity`, and they both are also `ABOUT`.

The approach of Warren (1978) to organize fine-grained semantic relations into a multi-level hierarchy has been followed by Nastase & Szpakowicz (2003) who designed an inventory of 30 fine-grained relations, grouped into five coarse-grained super-relations (the corresponding fine-grained relations are shown in parentheses): CAUSALITY (cause, effect, detraction, purpose), PARTICIPANT (agent, beneficiary, instrument, object_property, object, part, possessor, property, product, source, whole, stative), QUALITY (container, content, equative, material, measure, topic, type), SPATIAL (direction, location_at, location_from, location), and TEMPORALITY (frequency, time_at, time_through). For example, *exam anxiety* is classified as effect and therefore also as CAUSALITY, while *blue book* is property and therefore also PARTICIPANT.

This inventory of Nastase & Szpakowicz (2003) is an extension of the flat, non-hierarchical inventory of 20 relations created by Barker & Szpakowicz (1998). It is also a superset of the 19-relation inventory of Kim & Baldwin (2006).

There have been also some independently-developed inventories, which are not directly or indirectly derived from those of Levi (1978) and Warren (1978).

One such example is the inventory of Vanderwende (1994), which classifies noun-noun compounds according to questions (mostly *Wh-questions*), which in turn correspond to 13 semantic relations: Subject (*Who/what?*), Object (*Whom/what?*), Locative (*Where?*), Time (*When?*), Possessive (*Whose?*), Whole-Part (*What is it part of?*), Part-Whole (*What are its parts?*), Equative (*What kind of?*), Instrument (*How?*), Purpose (*What for?*), Material (*Made of what?*), Causes (*What does it cause?*), and Caused-by (*What causes it?*). For example, *alligator shoes* is Material and answers the question *Made of what?*.

Another example is the inventory of Girju & al. (2005), which consists of 21 abstract relations: POSSESSION, ATTRIBUTE-HOLDER, AGENT, TEMPORAL, PART-WHOLE, IS-A, CAUSE, MAKE/PRODUCE, INSTRUMENT, LOCATION/SPACE, PURPOSE, SOURCE, TOPIC, MANNER, MEANS, THEME, ACCOMPANIMENT, EXPERIENCER, RECIPIENT, MEASURE, and RESULT. It is a subset of the 35-relation inventory of Moldovan & al. (2004), originally designed for the semantic interpretation of noun phrases in general.

Finally, there has been some work on trying to compare and consolidate existing inventories. Tratz & Hovy (2010) compared the inventories of Levi (1978), Warren (1978), Vanderwende (1994), Barker & Szpakowicz (1998), Nastase & Szpakowicz (2003), and Girju & al. (2005) and found them to cover a largely overlapping semantic space but to partition this space differently. Thus, they proposed a new two-level inventory of 43 fine-grained relations, grouped into ten coarse-grained super-relations; Tratz & Hovy obtained this inventory through iterative crowd-sourcing by maximizing inter-annotator agreement.

While the above abstract relation inventories were designed with broad-coverage text analysis in mind, there have been also proposals tailored to a specific domain. For example, Rosario & al. (2002) defined 38 abstract relations to characterize the semantics of biomedical noun-noun compounds, which includes domain-specific relations such as Produce_Genetically, Person_Afflicted, Attribute_of_Clinical_Study, and Defect. For example, *hormone deficiency* is Defect, and *polyomavirus genome* is Produce_Genetically.

Another interesting proposal is that of Lauer (1995), who defined the problem of noun compound interpretation as predicting which among the following eight prepositions best paraphrases the target noun compound: `of`, `for`, `in`, `at`, `on`, `from`, `with`, and `about`. For example, *olive oil* is *oil from olives*, and *war story* is *story about war*. Lauer's approach is attractive since it is simple and allows direct extraction of prepositions from a large text corpus or from the Web (Lapata and Keller2004; Lapata and Keller2005). Moreover, prepositions are directly usable as paraphrases in NLP applications. The downside is that prepositions are polysemous: for example, `in`, `on`, and `at` can refer to both `LOCATION` and `TIME`, e.g., *mouse in the field* and *flight in the morning* express different relations, while *flight in the morning*, *flight at night* and *flight on Saturday* express the same relation. Similarly, *from* can correspond to several of Levi's relations (see Table 1), e.g., *tears from gas* is $\texttt{CAUSE}_1$, *death from drugs* is $\texttt{CAUSE}_2$, *cake from apples* is $\texttt{HAVE}_1$, *peel from lemon* is $\texttt{HAVE}_2$, *silk from worms* is $\texttt{MAKE}_1$, *ball from snow* is $\texttt{MAKE}_2$, *mouse from the fields* is `IN`, and *oil from olives* is `FROM`. Note also that some noun-noun compounds cannot be paraphrased with prepositions, e.g., *coach trainer*; in contrast, empirical work suggests that most noun-noun compounds do fall into one of Levi's RDPs.

The interpretation of noun compound in terms of abstract semantic relations drawn from a small fixed inventory is also potentially problematic for a number of reasons. First, it is unclear which relation inventory is best to use in general and for a particular task. Second, being both abstract and limited, such relations capture only part of the semantics, e.g., classifying *malaria mosquito* as `CAUSE` obscures the fact that mosquitos do not directly cause malaria, but just transmit it. Third, in many cases, multiple relations are possible, e.g., in Levi's theory, *sand dune* is interpretable as both `HAVE` and `BE`. Still, despite these issues, the use of abstract relation inventories remains the mainstream approach in computational linguistics.

In theoretical linguistics, however, the feasibility of the idea has been subject to much criticism. For example, after having performed a series of psychological experiments, Downing (1977) concluded that no fixed and limited set of relations could adequately characterize noun compound semantics, a position held earlier by Jespersen (1942). To see this, consider a noun compound like *plate length*, which can be interpreted in some context as *what your hair is when it drags in your food*. Obviously, such fine-grained novel interpretations cannot be covered by a finite fixed inventory of abstract semantic relations. This has led some computational linguists such as Hobbs & al.(1993) to assume that the relationship between the nouns in a noun-noun compound can be anything.

Some researchers have proposed a partial solution based on finer-grained, and even infinite, inventories, e.g., Finin (1980) used specific verbs such as *dissolved in* for *salt water*. While this still cannot handle examples such as *plate length*, it offers much richer representation.

Inspired by Finin (1980), Nakov & Hearst (2006), proposed that noun compound semantics is best expressible using paraphrases involving verbs and/or prepositions; this work was further extended in (Nakov2008b). For example, *bronze statue* is a statue that *is made of, is composed of, consists of, contains, is of, is, is handcrafted from, is dipped in, looks like* bronze.

Nakov & Hearst (2006) further proposed that selecting one such paraphrase is not enough to express the semantics of a noun compound and that multiple paraphrases are needed for a fine-grained representation. Finally, they observed that not all paraphrases are equally good (e.g., *is made of* is arguably better than *looks like* or *is dipped in* for MAKE), and thus proposed that the semantics of a noun compound should be expressed as a *distribution* over multiple possible paraphrases, e.g., *malaria mosquito* can be *carry (23)*, *spread (16)*, *cause (12)*, *transmit (9)*, etc. These verbs are fine-grained, directly usable as paraphrases, and using multiple of them for a noun compound approximates its semantics better.

It is easy to see that not only the semantics of individual noun compounds but also that of abstract relations such as MAKE can be represented in the same way, as a distribution over paraphrasing verbs and prepositions; this is an idea explored in (Nakov and Hearst2008). Note, however, that some noun compounds are paraphrasable by more specific verbs that do not necessarily support the target abstract relation. For example, *malaria mosquito*, which expresses CAUSE, can be paraphrased using verbs like *carry*, which do not imply direct causation. Similarly, while both *wrinkle treatment* and *migraine treatment* express TREATMENT-FOR-DISEASE, fine-grained differences can be revealed using verbs, e.g., *smooth* can paraphrase the former, but not the latter.

Finally, note that while verbs and prepositions are the most frequent ways to paraphrase a noun-noun compound, there are many other ways to express the explicit relationship between the two nouns, e.g., for *onion tears*, we could have: *tears from onions*, *tears due to cutting onion*, *tears induced when cutting onions*, *tears that onions induce*, *tears that come from chopping onions*, *tears that sometimes flow when onions are chopped*, *tears that raw onions give you*, etc. The exploration of such free paraphrases of noun-noun compounds was the focus of SemEval-2013 Task 4 (Hendrickx et al.2013).[15]

4.2 Semantic Interpretation

The semantic interpretation of noun compounds is complicated by their heterogeneous nature. Thus, it is often addressed using nonparametric instance-based classifiers like the k-nearest neighbor (kNN), which effectively reduce it to *measuring the relational similarity* between a testing and each of the training examples. The latter is studied in detail by Turney (2006b), who distinguishes between *attributional similarity* or correspondence between attributes, and *relational similarity* or correspondence between relations. Attributional similarity is interested in the similarity between two *words*, A and B. In contrast, relational similarity focuses on the relationship between two *pairs* of words, i.e., it asks how similar the relations A:B and C:D are. Measuring relational similarity is often done indirectly, and is modeled as a function of two instances of attributional similarity: (1) between A and C, and (2) between B and D.

[15] `http://www.cs.york.ac.uk/semeval-2013/task4/`

Going back to semantic relations,[16] there is a similar split between two general lines of research. The first one derives the noun compound semantics from the semantics of the nouns it is made of (Rosario and Hearst2001; Rosario et al.2002; Girju et al.2003; Moldovan et al.2004; Kim and Baldwin2005; Girju2006; Nastase et al.2006; Ó Séaghdha and Copestake2008; Tratz and Hovy2010), e.g., by generalizing them over a lexical hierarchy. This works well for relations like `Part-Whole`, which are more permanent and context-independent, e.g., *door-car*. The second line of research models the relationship between the nouns directly (Vanderwende1994; Lauer1995; Lapata2002; Turney and Littman2005; Kim and Baldwin2006; Nakov and Hearst2006; Turney2006b; Turney2006a; Nakov and Hearst2008; Butnariu and Veale2008), e.g., using suitable patterns that can connect them in a sentence. This is useful for context-dependent relations like `Cause-Effect`, which are dynamic and often episodic, e.g., *exam anxiety*. These two lines are rarely combined; two notable exceptions are (Séaghdha and Copestake2009) and (Nakov and Kozareva2011).

4.2.1 Attributional Approaches

A major advantage of attributional approaches is that they can make use of rich pre-existing resources such as WordNet, e.g., to generalize the nouns forming the noun compound. Below we present representative research in that direction.

The descent of hierarchy was proposed by Rosario & al. (2002), motivated by the assumption that head-modifier relations reflect the *qualia structure* (Pustejovsky1995) associated with the head. Under this interpretation, the meaning of the head determines what can be done to it, what it is made of, what it is a part of, and so on. For example, a *knife* can be in the following relations:

- Used-in: *kitchen knife*, *hunting knife*
- Made-of: *steel knife*, *plastic knife*
- Instrument-for: *carving knife*
- Used-on: *meat knife*, *putty knife*
- Used-by: *chef's knife*, *butcher's knife*

Some relations are specific to narrow noun classes, while others, more general, apply to wider classes. Building on this idea, Rosario & al. (2002) proposed a semi-supervised characterization of the relation between the nouns in a bioscience noun-noun compound, based on the semantic category in a lexical hierarchy to which each of the nouns belongs.

For example, all noun-noun compounds in which the first noun (the modifier) is classified under the A01 (*Body Regions*) sub-hierarchy and the second one (the head) falls under the A07 (*Cardiovascular System*) sub-hierarchy were hypothesized to express the same relation. Examples include *mesentery artery*, *finger capillary*, *leg vein*, and *forearm microcirculation*.

In contrast, noun-noun compounds whose constituent nouns are assigned to the A01-M01 (*Body Regions–Persons*) categories are ambiguous, and a distinction is needed between different kinds of persons:

[16] See (Nastase et al.2013) for a general overview on semantic relations.

- M01.643 (*Patients*), e.g., *eye outpatient*
- M01.526 (*Occupational Groups*), e.g., *eye physician*
- M01.898 (*Donors*), e.g., *eye donor*
- M01.150 (*Disabled Persons*), e.g., *arm amputee*

Thus, one needs to descend one level down the M01 hierarchy in order to find the right level of generalization so that all of the corresponding noun-noun compounds express the same relation.

The idea of the descent of hierarchy is appealing and the accuracy is very high (about 90%), but there are limitations. First, the classification is not fully automated; human annotators decide where to cut the hierarchy. Second, the coverage is limited by the lexical hierarchy, most likely to narrow domains. Third, problems are caused by lexical and relational ambiguities. Finally, the approach does not propose explicit names for the assigned relations.

Iterative semantic specialization (Girju et al.2003) is similar but automated and restricted to a single relation, e.g., Girju & al. (2003) experimented with `Part-Whole`. The training data, annotated with WordNet senses and containing both positive and negative examples, are first generalized going from each positive/negative example up the WordNet hierarchy, and then these generalizations are iteratively specialized whenever necessary, to make sure that they are not ambiguous with respect to the semantic relation assigned. A set of rules is then produced based on these examples through supervised learning.

Semantic scattering (Moldovan et al.2004) implemented an idea along the same lines. It relies on the training data to determine a boundary (essentially, a cut) in a hierarchy – in particular, WordNet's hypernym-hyponym hierarchy – such that the sense combinations in a noun compound which can be mapped onto this boundary are unambiguous with respect to the relation within the compound. Sense combinations found above this boundary may convey different relation types.

The interpretation of noun-noun compounds can be also done using the semantics of the head and of the modifier implicitly, i.e., without explicitly assigning them word senses. For example, Rosario & Hearst (2001) generalized noun compounds over the MeSH hierarchy at various levels of generality, without trying to disambiguate them; they then used these generalizations as features for predicting 18 relations between biomedical noun-noun compound in a supervised fashion. Similarly, Kim & Baldwin (2005) measured the similarity between two noun compounds as a linear interpolation of the WordNet similarity between the two heads and the two modifiers; then, they used this similarity in a kNN classifier.

Of course, nouns do not have to be generalized with respect to a hierarchy only; distributional information from a corpus might be equally useful. For example, Nastase & al. (2006), in addition to noun generalizations using WordNet, also used grammatical collocates extracted from a large corpus that appear with the head/modifier in a grammatical relation, such as subject, object, and prepositional complement. Similarly, Ó Séaghdha & Copestake (2008) measured similarity using collocates extracted from BNC and the Google Web 1T 5-gram corpus (Brants and Franz2006), which they used in distributional kernels.

Tratz & Hovy (2010) used another useful information source: morphological information such as noun prefixes and suffixes. Finally, Girju (2006) generalized arguments using cross-linguistic evidence from two bilingual sentence-aligned corpora: EUROPARL (Koehn2005) and CLUVI[17].

4.2.2 Relational Approaches

Relational approaches model the relationship between the nouns directly.

One way to do this is by using explicit paraphrases as a source of fine-grained semantic representation. We have already seen this in the work of Lauer (1995), who expressed noun compound semantics using eight prepositions, as well as in (Nakov and Hearst2006), who used a distribution over Web-derived verbs and prepositions. Butnariu & Veale (2008) also used paraphrasing verbs, but they extracted them as the intersection of possible head-verb and verb-modifier pairs. Using a distribution over verbs as a semantic interpretation was also carried out in a recent challenge: SemEval-2010 Task 9 (Butnariu et al.2009; Butnariu et al.2010).

Another way to model the relationship between the nouns in a noun compound is to use explicit paraphrases as features for predicting coarse-grained abstract relations. For example, Vanderwende (1994) associated verbs extracted from definitions in an online dictionary with abstract relations. Kim & Baldwin (2006) used verbs from the textual definitions of abstract semantic relations, augmented with similar verbs based on `WordNet::Similarity` (Pedersen et al.2004). Nakov & Hearst (2008) used Web-derived verbs, prepositions and coordinating conjunctions as features to predict Levi's RDPs.

Paraphrases as features to predict abstract relations do not have to be restricted to verbs, prepositions, conjunctions or any part of speech: just about any pattern is fine, even patterns containing placeholders. For example, Turney (2006a) mined from the Web relation patterns such as "*Y * causes X*" for `Cause` (e.g., *cold virus*) and "*Y in * early X*" for `Temporal` (e.g., *morning frost*). These patterns do not have to be mined; they could be also fixed. For example, Turney & Littman (2005) characterized the relationship between the nouns forming a compound using the Web frequencies for 128 fixed phrases like "*X* for *Y*" and "*Y* for *X*" instantiated from a fixed set of 64 joining terms such as *for*, *such as*, *not the*, *is **. These Web frequencies can be used for predicting abstract relations directly (Turney and Littman2005), or can be first mapped to vectors of lower dimensionality, e.g., using *Singular Value Decomposition* (Turney2006b).

A general problem with using paraphrases is that they suffer from data sparseness issues in the case of rare or rarely co-occurring nouns. This can be alleviated by combining relational features (e.g., verbs, prepositions and coordinating conjunctions) with attributional features (e.g., hypernyms and co-hyponyms of the nouns). The potential has already been demonstrated by Ó Séaghdha (2009), who used convolution kernels. A combination was also used in (Nakov and Kozareva2011).

[17] CLUVI – Linguistic Corpus of the University of Vigo – Parallel Corpus 2.1 – `http://sli.uvigo.es/CLUVI/`

5 Entailment

Below we discuss what subparts of a noun compound are entailed by the whole, and how this relates to syntax, semantics, and type of compound. We then show how interpreting noun compounds could help solve textual entailment problems.

5.1 Semantic Entailment

We have seen that for a right-headed endocentric compound n_1n_2, the relationship between the two nouns can be summarized as "n_1n_2 is a type/kind of n_2", e.g., for *lung cancer* we have that "*lung cancer* is a kind/type of *cancer*". We can interpret this paraphrase as a semantic entailment: that the compound n_1n_2 semantically entails n_2 in the sense that if we substitute n_2 in a statement containing n_1n_2, we will obtain a new statement that is entailed from the original one.

Of course, this does not hold for all noun-noun compounds, e.g., it fails for exocentric compounds (*birdbrain* is not a kind/type of *brain*; it is a kind of person), or for left-headed ones (*vitamin D* is not a kind/type of *D*; it is a kind of *vitamin*).

For a three-word right-headed endocentric noun compound $n_1n_2n_3$, it can be expected that the assertion "$n_1n_2n_3$ is a type/kind of n_3" should be true, e.g., "*lung cancer drug* is a kind/type of *drug*". If the compound is left-bracketed $(n_1n_2)n_3$, it can be further expected that the assertion "$n_1n_2n_3$ is a type/kind of n_2n_3" would hold as well, e.g., "*lung cancer drug* is a kind/type of *cancer drug*". If the compound is right-bracketed $n_1(n_2n_3)$, then it can be expected that the following two assertions would be true "$n_1n_2n_3$ is a type/kind of n_2n_3" and "$n_1n_2n_3$ is a type/kind of n_1n_3", e.g., "*oxydant wrinkle treatment* is a kind/type of *wrinkle treatment*" and "*oxydant wrinkle treatment* is a kind/type of *oxydant treatment*".

In other words, an endocentric right-headed nonlexicalized three-word noun compound $n_1n_2n_3$ is expected to entail n_3 and n_2n_3, but not n_1n_3, if left-bracketed, and all three n_3, n_2n_3, and n_1n_3, if right-bracketed. Therefore, noun compound bracketing could be used to predict semantic entailment. Unfortunately, it is hard for a computer program to distinguish left-headed vs. right-headed compound, or left-bracketed vs. right-bracketed compound; it is even harder to distinguish between a lexicalized and a nonlexicalized compound, the boundary between which is not clear-cut. The matter is further complicated due to idiomacity, lexicalization, transparency, world knowledge, nominalization, left headedness, and metonymy.

Table 3 shows sample three-word noun compounds, the corresponding bracketing and whether n_1n_3 and n_2n_3 are entailed or not. The entailments that are not predicted by the bracketing are shown in bold; we can easily explain some of them:

1. **Idiomacity**. Idiomacity can interfere with semantic entailment by blocking the extraction of the head from a pair of bracketed words. This is the case with [*butter ball*] *guy*, where the idiomacity of *butter ball* ("fat person") does not allow the extraction of *ball* (i.e., here *butter ball* is not a kind of *ball*), and thus prevents the entailment of *ball guy*. The same applies to [*lab lit*] *book*, where *lab lit*, a literary genre, is idiomatic and blocks the extraction of *lit* (i.e., *lab lit* is not a kind of *lit*), and ultimately the entailment of *lit book*.

$n_1n_2n_3$	n_1n_3	"⇒"	n_2n_3	"⇒"
Left-bracketed				
army ant behavior	*army behavior*	no	*ant behavior*	yes
lung cancer doctor	*lung doctor*	**YES**	*cancer doctor*	yes
lung cancer patient	*lung patient*	**YES**	*cancer patient*	yes
lung cancer survivor	*lung survivor*	no	*cancer survivor*	yes
alien invasion book	*alien book*	**YES**	*invasion book*	yes
butter ball guy	*butter guy*	no	*ball guy*	**NO**
lab lit book	*lab book*	no	*lit book*	**NO**
I novel writer	*I writer*	no	*novel writer*	yes
US army forces	*US forces*	**YES**	*army forces*	yes
vitamin D deficiency	*vitamin deficiency*	**YES**	*D deficiency*	**NO**
planet Earth survival	*planet survival*	**YES**	*Earth survival*	yes
Right-bracketed				
oxydant wrinkle treatment	*oxydant treatment*	yes	*wrinkle treatment*	yes
vanilla ice cream	*vanilla cream*	**NO**	*ice cream*	yes

Table 3. ***Semantic entailment and noun compound bracketing.*** *The entailments that are not predicted by the bracketing are shown in bold.*

2. **Lexicalization**. Another explanation for [*butter ball*] *guy* and [*lab lit*] *book* is lexicalization since idiomatic compounds are necessarily lexicalized. Lexicalization also explains the case of *vanilla* [*ice cream*], where *ice cream* does not allow the extraction of *cream*, thus blocking the entailment of *vanilla cream*.
3. **Transparency**. Lexicalization cannot explain all examples in the table, most of which include subparts that are at least partially lexicalized. For example, [*army ant*] *behavior* includes the lexicalized compound *army ant*; still, this does not prevent the extraction of the head *ant* and thus the entailment of *ant behavior*. This extraction is possible since *army ant* is transparent. The same argument applies to compounds that contain even more highly lexicalized subparts, e.g., *novel* can be extracted from *I novel*, a kind of Japanese novel, so that *novel writer* can be entailed from [*I novel*] *writer*.
4. **World knowledge**. The entailment of *lung doctor* from *lung cancer doctor* cannot be explained by the relatively low degree of lexicalization of *lung cancer* alone. The noun compound is left bracketed and this entailment is not predicted by the bracketing. Here, the semantics of the individual nouns and world knowledge come into play: *doctor* is likely to be modified by organs.
5. **Nominalization**. Two problematic noun compounds are *lung cancer patient* and *lung cancer survivor*. While *patient* and *survivor* play similar role in the context of these compounds, *lung patient* is entailed, but *lung survivor* is not.

Here *survivor* is a nominalization of the verb *survive*, which can easily take *cancer*, and therefore also *lung cancer*, but not *lung*, as a direct object. A similar argument about nominalization applies to [*Alien invasion*] *book* and [*planet Earth*] *survival*.

6. **Left headedness**. The actual entailments are completely reversed compared to what the bracketing predicts for *vitamin D deficiency*, which contains the left-headed noun compound *vitamin D* as a modifier.
7. **Metonymy**. Due to metonymy, in the case of *US army forces*, all possible subsequences are entailed: not only *US forces*, *army forces* and *forces*, but also *US army*, *army* and, in some contexts, even just *US*.

5.2 Application to Textual Entailment

In this section, we present some examples involving noun compounds from the development dataset of the Second Pascal Recognizing Textual Entailment (RTE2) Challenge.[18] We have chosen to illustrate the importance of understanding the syntax and semantics of noun compounds on this particular task since it requires generic semantic inference that could potentially help many natural language processing applications, including Question Answering (QA), Information Retrieval (IR), Information Extraction (IE), and (multi-)document summarization (SUM).

Given two textual fragments, a text T and a hypothesys H, the goal is to recognize whether the meaning of H is entailed (can be inferred) from T:

"*We say that T entails H if, typically, a human reading T would infer that H is most likely true. This somewhat informal definition is based on (and assumes) common human understanding of language as well as common background knowledge.*" (RTE2 task definition)

Below we give examples illustrating different kinds of phrase variability involving noun compounds that a successful inference engine for RTE might need to solve. Of course, recognizing that two phrases are variations of each other does not solve the RTE problem automatically (e.g., below we have both positive and negative examples with respect to RTE), but it could help the overall decision by allowing an RTE engine to focus on the remaining words rather than on the surface variability between noun compounds and their paraphrases/subparts.

5.2.1 Prepositional Paraphrase

Here we have a positive example from IR, where a noun compound has to be matched to a prepositional paraphrase:

"*paper costs*" ⇒ "*cost of paper*"

[18] `http://www.pascal-network.org/Challenges/RTE2`

```
<pair id="503" entailment="YES" task="IR">
<t>Newspapers choke on rising paper costs and falling revenue.</t>
<h>The cost of paper is rising.</h>
</pair>
```

5.2.2 Verbal & Prepositional Paraphrases

Here we have a positive example from QA, which requires matching two different paraphrases, one verbal and one prepositional, of the same underlying three-word noun compound *WTO Geneva headquarters*:

"*Geneva headquarters of the WTO*" ⇒ "*WTO headquarters are located in Geneva*"

```
<pair id="284" entailment="YES" task="QA">
<t>While preliminary work goes on at the Geneva headquarters of the WTO,
with members providing input, key decisions are taken at the ministerial
meetings.</t>
<h>The WTO headquarters are located in Geneva.</h>
</pair>
```

5.2.3 Bracketing

Here is a positive example from Summarization, where the hypothesis extracts the modifier of a three-word noun compound, which is consistent with a left bracketing:

"[*breast cancer*] *patients*" ⇒ "[*breast cancer*]"

```
<pair id="177" entailment="YES" task="SUM">
<t>Herceptin was already approved to treat the sickest
breast cancer patients, and the company said, Monday,
it will discuss with federal regulators the possibility
of prescribing the drug for more breast cancer patients.</t>
<h>Herceptin can be used to treat breast cancer.</h>
</pair>
```

5.2.4 Bracketing & Paraphrase

The following negative example from IR involves a noun compound paraphrased prepositionally in a way that is consistent with left bracketing:

"[*ivory trade*] *ban*" ⇒ "*ban on* [*ivory trade*]"

```
<pair id="117" entailment="NO" task="IR">
<t>The release of its report led to calls for a complete ivory trade ban,
and at the seventh conference in 1989, the African Elephant was moved to
appendix one of the treaty.</t>
<h>The ban on ivory trade has been effective in protecting the elephant
from extinction.</h>
</pair>
```

5.2.5 Synonymy

Here is another IR example, this time positive, which involves two noun compounds with the same modifiers and with synonymous heads, which express the same implicit semantic relation:

"*marine plants*" ⇒ "*marine vegetation*"

```
<pair id="65" entailment="YES" task="IR">
<t>A number of marine plants are harvested commercially
in Nova Scotia.</t>
<h>Marine vegetation is harvested.</h>
</pair>
```

5.2.6 Hyponymy & Prepositional Paraphrase

The following positive example from IR involves a prepositional paraphrase and a noun compound, where the modifiers are the same and the heads are in a hyponymy relation, i.e., *marijuana* is a kind of *drug*:

"*legalization of marijuana*" ⇒ "*drug legalization*"

We could also see this as involving longer paraphrases for the noun compounds *marijuana legalization benefits* and *drug legalization benefits*:

"*benefits in the legalization of marijuana*" ⇒ "*drug legalization has benefits*".

```
<pair id="363" entailment="YES" task="IR">
<t>One economic study will not be the basis of Canada's public policy
 decisions, but Easton's research does conclusively show that there are
economic benefits in the legalization of marijuana.</t>
<h>Drug legalization has benefits.</h>
</pair>
```

5.2.7 Prepositional Paraphrase, Nominalization & Synonymy

The following example is yet another positive example for IR, which involves a combination of several phenomena. Here the verb *enrich* was substitued by its synonym *enhance*, which in turn was nominalized. On the other hand, the verbs *enrich* and *feed* are good paraphrases for the noun-noun compound *soil enhancers*, e.g., "*enhancers that enrich/feed the soil*", and *fertilizer* is a synonym of *enhancer*:

"*enriches and feeds the soil*" ⇒ "*soil enhancers*"

```
<pair id="192" entailment="YES" task="IR">
<t>Organic fertilizer slowly enriches and feeds the soil. Fast acting
synthetic fertilizers harm soil life.</t>
<h>Organic fertilizers are used as soil enhancers.</h>
</pair>
```

5.2.8 Nominalization, Bracketing & Synonymy

The negative example from Summarization below is another illustration of complex interactions. Here the modificational past participle *damaged* becomes the new noun compound head: *injury*. One possible interpretation is that *damaged*, which is a form of the verb *to damage*, has been substituted by the synonymous verb *to injure*, which has been nominalized and has become the new noun compound head:

"*damaged* [*spinal cord*]" ⇒ "[*spinal cord*] *injury*"

Note, however, that the verb *to damage* does not have to be analyzed as a synonym of *to injure*; it is also a good candidate for a verbal paraphrase of the noun compound *spinal cord injury*, e.g., as "*injury which damaged the spinal cord*".

```
<pair id="261" entailment="NO" task="SUM">
<t>The new work went an extra step, suggesting that the connections that
the stem cells form to help bridge the damaged spinal cord, are key to
recovery.</t>
<h>The experiment, reported Monday, isn't the first to show that stem
cells offer tantalizing hope for spinal cord injury.</h>
</pair>
```

6 Discussion

We have seen above that solving many textual entailment problems might require understanding the syntax and/or the semantics of the involved noun compounds. Some of the examples needed to check whether a particular verb or preposition is acceptable as a paraphrase for a given noun compound, others called for deciding whether semantic entailment would be possible after dropping one or more words from a noun compound. Additional factors come into play, e.g., inflectional and derivational morphological alternations and lexical relations such as synonymy and hyponymy, as well as complex combinations thereof.

An important observation to make is that noun compound interpretation techniques can extend to more complex expressions. For example, checking whether "*Geneva headquarters of the WTO*" could entail "*WTO headquarters are located in Geneva*" can be performed by understanding the semantics of the noun compound "*WTO Geneva headquarters*" in terms of prepositional and verbal paraphrases.

Similarly, even without knowing that *plant* and *vegetation* are synonyms, we can still establish that *marine plants* textually entails *marine vegetation* since the implicit semantic head-modifier relations for these noun compounds are very similar, e.g., can be paraphrased using the same verbs and/or prepositions.

We should note that the task solved by RTE is probably more complex than many real-world application would require.[19] For example, an IR engine would benefit directly from being able to recognize that "*Geneva headquarters of the WTO*", "*WTO headquarters are located in Geneva*" and "*WTO Geneva headquarters*" are all paraphrases, without the need for solving the RTE problem for full sentences.

Being able to generate such paraphrases by turning a noun compound into an explicit paraphrase and vice versa, has been shown beneficial (Nakov2008a) for statistical machine translation (SMT). For example, suppose that the phrase *oil price hikes* can be paraphrased as *hikes in oil prices* and *hikes in the prices of oil*. If an SMT system knows that *oil price hikes* can be translated into Spanish as *alzas en los precios del petróleo*,[20] then it can also use the same fluent Spanish translation for the two additional paraphrases.

One notable feature of the RTE task is that noun compounds are put into a sentential context, which is more realistic than out-of-context interpretation since the semantics of many noun compounds is context-dependent (Meyer1993), e.g., *museum book* can mean *a book about a museum*, *a book on display in a museum*, *a book published by a museum*, *a book showing a museum on its cover page*, *a book bought in a museum*, etc. More importantly, compounds could get completely novel interpretations in a certain context, e.g., *museum book* could mean *a book that I am planning to take with me when I go to the museum tomorrow*; such novel interpretations are currently out of reach for computational approaches.

Context-dependency is generally somewhat less of an issue for the semantic relations that hold between the nouns forming a noun compound compared to general semantic relations between nouns. This is partly due to the two nouns mutually disambiguating themselves, e.g., in the interpretations of *museum book* above, *book* had the meaning of either (a) physical object or (b) information source, but not of (c) the number of tricks a cardplayer or side must win before any trick can have scoring value. It is also due to most noun compounds preferring a particular interpretation, e.g., for *museum book* this would be *a book about a museum*.

[19] This is similar to other evaluation campaigns such as TREC, which have asked for answering unrealistically long full-sentence queries, while a typical Web query is only 1-3 words long (Spink et al.2001).

[20] This is hard to generate word-for-word for the English input because of the need to also generate the connecting prepositions and determiners *en los* and *del*, but it could be extracted as a readily available phrase pair from a parallel sentence-aligned bi-text as part of the process of training a phrase-based SMT system (Koehn et al.2003).

The general preference for one particular default interpretation is reflected in the available noun compound datasets, e.g., those of Levi (1978), Warren (1978), Vanderwende (1994), Lauer (1995), Barker & Szpakowicz (1998), Nastase & Szpakowicz (2003), Girju & al. (2005), Kim & Baldwin (2006), Diarmuid Ó Séaghdha (2007), and Tratz & Hovy (2010), none of which provides context for the interpretation.

This tradition of out-of-context interpretation was also continued by evaluation campaigns on noun compound interpretation such as SemEval-2010 Task 9 (Butnariu et al.2009; Butnariu et al.2010), which asks for semantic interpretation using paraphrasing verbs as opposed to using abstract relations. It assumed that candidate paraphrasing verbs and prepositions have already been identified by some hypothetical system, and asked the task participants to rank a long list of such candidates for a given noun-noun compound by relevance in decreasing order. For example, *cause* and *spread* are good paraphrases for *malaria mosquito*, and thus should be ranked high, while *be made up of* is bad and thus should be ranked low.

This is in contrast to related tasks such as SemEval-2007 Task 4 (Girju et al.2007; Girju et al.2009) and SemEval-2010 Task 8 (Hendrickx et al.2009; Hendrickx et al.2010), which asked for the classification of semantic relations between pairs of nominals in the context of a sentence. Context plays an important role in these tasks, e.g., while the noun compound *apple basket* can be only interpreted as `Content-Container` according to the inventory of semantic relations in SemEval-2010 Task 8, there are three possible context-dependent relations in that inventory between *apple* and *basket* when they do not form a compound, as the following examples show (illustrative made-up examples; not present in the task datasets):

`Entity-Destination`: "*A girl entered the room, put three red <e1>apples</e1> in the <e2>basket</e2>, and left.*"

`Entity-Origin`: "*A few minutes later, a boy came along, took the biggest <e1>apple</e1> from the <e2>basket</e2>, and ate it.*"

`Content-Container`: "*When the girl returned, she found that there were only two <e1>apples</e1> left in the <e2>basket</e2>.*"

We should note, however, that the datasets of SemEval-2007 Task 4 and SemEval-2010 Task 8 include some examples (8.2% and 5.1%, respectively, as Tables 4 and 5 show) that do ask for the interpretation of noun compounds. However, most of these noun compounds have their default interpretation and thus do not really need the sentential context, e.g. (this time, actual examples from the dataset):

`Entity-Origin`: "*<e1>Coconut</e1> <e2>oil</e2> is extracted from the kernel or meat of matured coconut harvested from the coconut palm (Cocos nucifera).*"

`Product-Producer`: "*The <e1>police</e1> <e2>report</e2> has cast a spotlight on America's self-help industry.*"

Relation	Example	Statistics	
Cause-Effect	*drinking problems*	4 / 220	1.8%
Instrument-Agency	*crane operator*	7 / 218	3.2%
Product-Producer	*mustard company*	35 / 233	15.0%
Origin-Entity	*cane sugar*	38 / 221	17.2%
Theme-Tool	*email message*	10 / 211	4.7%
Part-Whole	*apple seed*	32 / 212	15.1%
Content-Container	—	0 / 214	0.0%
		126 / 1529	8.2%

Table 4. *The noun-noun compounds in the SemEval-2007 Task 4 datasets.*

Relation	Example	Statistics	
Cause-Effect	*fatigue corrosion*	18 / 1331	1.4%
Instrument-Agency	*phone operator*	21 / 660	3.2%
Product-Producer	*coffee machine*	78 / 948	8.2%
Content-Container	*tea bags*	9 / 732	1.2%
Entity-Origin	*raspberry syrup*	153 / 974	15.7%
Entity-Destination	—	0 / 1137	0.0%
Component-Whole	*umbrella frames*	128 / 1253	10.2%
Member-Collection	*student association*	20 / 923	2.2%
Message-Topic	*crime fiction*	6 / 895	0.7%
Other	*diamond colliers*	111 / 1864	6.0%
		544 / 10717	5.1%

Table 5. *The noun-noun compounds in the SemEval-2010 Task 8 datasets.*

`Component-Whole`: "*To stand up to the wind, <e_1>umbrella</e_1> <e_2>frames</e_2> are strong yet flexible.*"

`Cause-Effect`: "*<e_1>Fatigue</e_1> <e_2>corrosion</e_2> and stress corrosion are similar, as both are caused by external stresses applied to the pipe and occur inside of the pipe.*"

`Other`: "*Tissue damage and <e_1>electrode</e_1> <e_2>corrosion</e_2> are both associated with high charge density stimulation.*"

The datasets of SemEval-2007 Task 4 and SemEval-2010 Task 8 are not the only ones asking for noun compound interpretation in context. There is also a dataset inspired by RTE, which asks specifically for resolving entailments between a paraphrase and a noun compound (Nakov2008c); here is a positive example:

```
T: I have friends that are organizing to get more <e2>professors</e2>
that are <e1>women</e> and educate women to make specific choices on where
to get jobs.
H: I have friends that are organizing to get more <e1>women</e1>
<e2>professors</e2> and educate women to make specific choices on where
to get jobs.
```

And here is a negative example, where a bad paraphrasing verb is used in the first sentence:

```
T: As McMillan collected, she also quietly gave, donating millions of
dollars to create scholarships and fellowships for black Harvard Medical
School students, African filmmakers, and MIT <e2>professors</e2> who study
<e1>women</e1> in the developing world.
H: As McMillan collected, she also quietly gave, donating millions
of dollars to create scholarships and fellowships for black Harvard
Medical School students, African filmmakers, and MIT <e1>women</e1>
<e2>professors</e2> in the developing world.
```

This dataset is probably too small for practical uses, but it points to an interesting research direction with many potential practical applications, e.g., such sentence-level paraphrases have already been shown useful in SMT (Nakov2008a).

7 Conclusion

We have presented a brief overview of noun compounds and their syntax and semantics from both theoretical and computational linguistics viewpoint with an emphasis on the latter. We have also shown how understanding noun compound syntax and semantics could potentially help solve textual entailment problems, both semantic entailment and RTE, which would be potentially useful for a number of NLP applications, and which we believe to be an important direction for future research.

Another important, but surprisingly under-explored, research direction we discussed is noun compound interpretation in context. While the importance of context and pragmatic considerations for noun compound interpretation are well understood in theoretical linguistics (Jespersen1942; Downing1977; Spärck Jones1983; Meyer1993), computational linguists have so far largely ignored context dependency,[21] even though they have modeled it for semantic relations in general. Facilitating research in this direction would require suitable datasets.

Exploring the relationship between fine-grained paraphrases, which are typically extracted in an unsupervised fashion, and abstract relations, which are used primarily with supervised models, is another promising research direction. This would be facilitated by datasets that have both kinds of annotations, e.g., SemEval-2010 Task 9 (Butnariu et al.2009; Butnariu et al.2010) provides verbal and prepositional paraphrases for Levi's examples, for which RDPs are also available; such datasets can be also automatically bootstrapped from the Web (Kim and Nakov2011).

[21] The work of Hobbs & al. (1993) is a notable exception.

More generally, there is a need for better understanding of the limitations and the advantages of different representation schemes; while there are some initial attempts in that direction (Tratz and Hovy2010; Kim and Nakov2011), much deeper analysis is necessary, which would compare different schemes in terms of expressiveness, potential for theoretical insight, usability for practical applications, etc.

Improved models for noun compound interpretation would require combining multiple knowledge sources (Tratz and Hovy2010), e.g., relational and attributional features (Séaghdha and Copestake2009; Nakov and Kozareva2011), Web-scale statistics extracted dynamically as needed (Nakov and Hearst2008) or from static resources such as the Google Web 1T 5-gram corpus (Butnariu and Veale2008), linguistically-motivated features and paraphrases (Nakov and Hearst2005a), pre-existing lexical resources such as WordNet (Kim and Baldwin2005) or Roget's thesaurus (Lauer1995), cross-linguistic evidence (Girju2007), etc.

Moreover, in view of the recent advances in modeling distributional compositionality for word combinations (Baroni and Zamparelli2010; Mitchell and Lapata2010; Socher et al.2012), we believe that this is another interesting research direction, which could offer important insights for noun compound interpretation.

Last but not least, there is a need for more applications. *Fine-grained paraphrases* of noun compounds in terms of verbs and prepositions have already been shown beneficial, e.g., for statistical machine translation, where they were used directly (Nakov2008a), or for predicting abstract relations from a small fixed inventory, with paraphrases used as features (Kim and Baldwin2006; Nakov and Hearst2008); we have also suggested potential applications to information retrieval, question answering, and textual entailment. However, the usability of *abstract* semantic relations for practical applications still remains to be proven in practice.

References

John Algeo, editor. 1991. *Fifty Years Among the New Words.* Cambridge University Press, Cambridge.

Timothy Baldwin and Takaaki Tanaka. 2004. Translation by machine of compound nominals: Getting it right. In *Proceedings of the ACL 2004 Workshop on Multiword Expressions: Integrating Processing*, MWE '04, pages 24–31, Barcelona, Spain.

Ken Barker and Stan Szpakowicz. 1998. Semi-automatic recognition of noun modifier relationships. In *Proceedings of the 17th International Conference on Computational Linguistics*, ICCL '98, pages 96–102, Chicago, IL.

Marco Baroni and Roberto Zamparelli. 2010. Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 1183–1193, Cambridge, MA.

Laurie Bauer. 1983. *English Word-Formation.* Cambridge University Press, Cambridge.

Laurie Bauer. 2006. Compound. *Linguistics and Philosophy*, 17:329–342.

Geert Booij. 2005. *The Grammar Of Words: An Introduction to Linguistic Morphology.* Oxford linguistics. Oxford University Press.

Thorsten Brants and Alex Franz. 2006. Web 1T 5-gram corpus version 1.1. Technical report, Linguistic Data Consortium, Philadelphia, PA.

Cristina Butnariu and Tony Veale. 2008. A concept-centered approach to noun-compound interpretation. In *Proceedings of the 22nd International Conference on Computational Linguistics*, COLING '08, pages 81–88, Machester, UK.

Cristina Butnariu, Su Nam Kim, Preslav Nakov, Diarmuid Ó Séaghdha, Stan Szpakowicz, and Tony Veale. 2009. SemEval-2010 task 9: The interpretation of noun compounds using paraphrasing verbs and prepositions. In *Proceedings of the NAACL-HLT-09 Workshop on Semantic Evaluations: Recent Achievements and Future Directions*, SEW '09, pages 100–105, Boulder, CO.

Cristina Butnariu, Su Nam Kim, Preslav Nakov, Diarmuid Ó Séaghdha, Stan Szpakowicz, and Tony Veale. 2010. SemEval-2 task 9: The interpretation of noun compounds using paraphrasing verbs and prepositions. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, pages 39–44, Uppsala, Sweden.

Noam Chomsky and Morris Halle. 1968. *Sound Pattern of English*. MIT Press, Cambridge, MA.

Noam Chomsky, Morris Halle, and Fred Lukoff. 1956. On accent and juncture in English. *For Roman Jakobson: Essays on the occasion of his sixtieth birthday*, pages 65–80.

Anna Maria Di Sciullo and Edwin Williams. 1987. *On the Definition of Word*. MIT Press, Cambridge, MA.

Pamela Downing. 1977. On the creation and use of English compound nouns. *Language*, 53(4):810–842.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Timothy Finin. 1980. *The Semantic Interpretation of Compound Nominals*. Ph.D. thesis, University of Illinois, Urbana, IL.

Roxana Girju, Adriana Badulescu, and Dan Moldovan. 2003. Learning semantic constraints for the automatic discovery of part-whole relations. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, NAACL '03, pages 1–8, Edmonton, Canada.

Roxana Girju, Dan Moldovan, Marta Tatu, and Daniel Antohe. 2005. On the semantics of noun compounds. *Journal of Computer Speech and Language - Special Issue on Multiword Expressions*, 4(19):479–496.

Roxana Girju, Preslav Nakov, Vivi Nastase, Stan Szpakowicz, Peter Turney, and Deniz Yuret. 2007. SemEval-2007 task 04: Classification of semantic relations between nominals. In *Proceedings of the Fourth International Workshop on Semantic Evaluations*, SemEval '07, pages 13–18, Prague, Czech Republic.

Roxana Girju, Preslav Nakov, Vivi Nastase, Stan Szpakowicz, Peter Turney, and Deniz Yuret. 2009. Classification of semantic relations between nominals. *Language Resources and Evaluation*, 43(2):105–121.

Roxana Girju. 2006. Out-of-context noun phrase semantic interpretation with cross-linguistic evidence. In *Proceedings of the 15th ACM international conference on Information and knowledge management*, CIKM '06, pages 268–276, Arlington, Virginia.

Roxana Girju. 2007. Experiments with an annotation scheme for a knowledge-rich noun phrase interpretation system. In *Proceedings of the Linguistic Annotation Workshop*, LAW '07, pages 168–175, Prague, Czech Republic.

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2009. SemEval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions*, SEW '09, pages 94–99, Boulder, CO.

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2010. SemEval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, pages 33–38, Uppsala, Sweden.

Iris Hendrickx, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Stan Szpakowicz, and Tony Veale. 2013. SemEval-2013 task 4: Free paraphrases of noun compounds.

In *Proceedings of the International Workshop on Semantic Evaluation*, SemEval '13, Atlanta, Georgia.

Jerry R. Hobbs, Mark E. Stickel, Douglas E. Appelt, and Paul Martin. 1993. Interpretation as abduction. *Artificial Intelligence*, 63(1-2):69–142.

Rodney Huddleston and Geoffrey Pullum. 2002. *The Cambridge Grammar of the English Language*. Cambridge University Press.

Ray Jackendoff. 1975. Morphological and semantic regularities in the lexicon. *Language*, 51:639–671.

Otto Jespersen. 1942. *A Modern English Grammar on Historical Principles Part VI: Morphology*. Ejaar Munksgaard, Copenhagen.

Su Nam Kim and Timothy Baldwin. 2005. Automatic interpretation of compound nouns using WordNet similarity. In *Proceedings of 2nd International Joint Conference on Natural Language Processing*, IJCNLP '05, pages 945–956, Jeju, Korea.

Su Nam Kim and Timothy Baldwin. 2006. Interpreting semantic relations in noun compounds via verb semantics. In *Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics and 21st International Conference on Computational Linguistics*, COLING-ACL '06, pages 491–498, Sydney, Australia.

Su Nam Kim and Preslav Nakov. 2011. Large-scale noun compound interpretation using bootstrapping and the web as a corpus. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 648–658, Edinburgh, UK.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, HLT-NAACL '03, pages 48–54, Edmonton, Canada.

Philipp Koehn. 2005. Europarl: A parallel corpus for evaluation of machine translation. In *Proceedings of the X MT Summit*, pages 79–86, Phuket, Thailand.

Henry Kucera and Nelson Francis. 1967. *Computational Analysis of Present-Day American English*. Brown University Press, Providence, RI.

Robert D. Ladd. 1984. English compound stress. In Dafydd Gibbon and Helmut Richter, editors, *Intonation, Accent and Rhythm: Studies in Discourse Phonology*. W de Gruyter, Berlin.

Mirella Lapata and Frank Keller. 2004. The Web as a baseline: Evaluating the performance of unsupervised Web-based models for a range of NLP tasks. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, HLT-NAACL '04, pages 121–128, Boston, MA.

Mirella Lapata and Frank Keller. 2005. Web-based models for natural language processing. *ACM Trans. Speech Lang. Process.*, 2(1):3.

Maria Lapata. 2002. The disambiguation of nominalizations. *Computational Linguistics*, 28(3):357–388.

Mark Lauer. 1995. *Designing Statistical Language Learners: Experiments on Noun Compounds*. Ph.D. thesis, Department of Computing, Macquarie University, Australia.

Judith Levi. 1978. *The Syntax and Semantics of Complex Nominals.* Academic Press, New York.

Mark Liberman and Richard Sproat. 1992. The stress and structure of modified noun phrases in English. In Ivan A. Sag and Anna Szabolcsi, editors, *Lexical Matters*, pages 131–181. CSLI Publications, Stanford.

Rochelle Lieber and Pavol Stekauer, editors. 2009. *The Oxford Handbook of Compounding*. Oxford Handbooks in Linguistics. OUP Oxford.

Mitchell Marcus. 1980. *A Theory of Syntactic Recognition for Natural Language*. MIT Press, Cambridge, MA.

Ralf Meyer. 1993. *Compound Comprehension in Isolation and in Context: The Contribution of Conceptual and Discourse Knowledge to the Comprehension of German Novel Noun-Noun Compounds*. Linguistische Arbeiten 299. Niemeyer, Tübingen.

Jeff Mitchell and Mirella Lapata. 2010. Composition in distributional models of semantics. *Cognitive Science*, 34(8):1388–1429.

Dan Moldovan, Adriana Badulescu, Marta Tatu, Daniel Antohe, and Roxana Girju. 2004. Models for the semantic classification of noun phrases. In *Proceedings of the HLT-NAACL Workshop on Computational Lexical Semantics*, pages 60–67, Boston, MA.

Preslav Nakov and Marti Hearst. 2005a. Search engine statistics beyond the n-gram: Application to noun compound bracketing. In *Proceedings of the Ninth Conference on Computational Natural Language Learning*, CoNLL '05, pages 17–24, Ann Arbor, MI.

Preslav Nakov and Marti Hearst. 2005b. A study of using search engine page hits as a proxy for n-gram frequencies. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, RANLP '05, pages 347–353, Borovets, Bulgaria.

Preslav Nakov and Marti Hearst. 2005c. Using the web as an implicit training set: application to structural ambiguity resolution. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT-EMNLP '05, pages 835–842, Vancouver, Canada.

Preslav Nakov and Marti Hearst. 2006. Using verbs to characterize noun-noun relations. In Jerome Euzenat and John Domingue, editors, *AIMSA*, volume 4183 of *Lecture Notes in Computer Science*, pages 233–244. Springer.

Preslav Nakov and Marti Hearst. 2008. Solving relational similarity problems using the web as a corpus. In *Proceedings of the 46th Annual Meeting on Association for Computational Linguistics*, ACL '08, pages 452–460, Columbus, OH.

Preslav Nakov and Zornitsa Kozareva. 2011. Combining relational and attributional similarity for semantic relation classification. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, RANLP '11, pages 323–330, Hissar, Bulgaria.

Preslav Nakov. 2007. *Using the Web as an Implicit Training Set: Application to Noun Compound Syntax and Semantics*. Ph.D. thesis, EECS Department, University of California, Berkeley, UCB/EECS-2007-173.

Preslav Nakov. 2008a. Improved statistical machine translation using monolingual paraphrases. In *Proceedings of the European Conference on Artificial Intelligence*, ECAI '08, pages 338–342, Patras, Greece.

Preslav Nakov. 2008b. Noun compound interpretation using paraphrasing verbs: Feasibility study. In *Proceedings of the 13th international conference on Artificial Intelligence: Methodology, Systems, and Applications*, AIMSA '08, pages 103–117, Varna, Bulgaria.

Preslav Nakov. 2008c. Paraphrasing verbs for noun compound interpretation. In *Proceedings of the LREC'08 Workshop: Towards a Shared Task for Multiword Expressions*, MWE '08, pages 46–49, Marrakech, Morocco.

Vivi Nastase and Stan Szpakowicz. 2003. Exploring noun-modifier semantic relations. In *Proceedings of the Fifth International Workshop on Computational Semantics*, IWCS '03, pages 285–301, Tilburg, Holland.

Vivi Nastase, Jelber Sayyad-Shirabad, Marina Sokolova, and Stan Szpakowicz. 2006. Learning noun-modifier semantic relations with corpus-based and WordNet-based features. In *Proceedings of the 21st National Conference on Artificial Intelligence*, pages 781–787, Boston, MA.

Vivi Nastase, Preslav Nakov, Diarmuid Ó Séaghdha, and Stan Szpakowicz. 2013. *Semantic Relations between Nominals*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Diarmuid Ó Séaghdha and Ann Copestake. 2008. Semantic classification with distributional kernels. In *Proceedings of the 22nd International Conference on Computational Linguistics*, COLING '08, pages 649–656, Manchester, UK.

Diarmuid Ó Séaghdha. 2007. Designing and evaluating a semantic annotation scheme for compound nouns. In *Proceedings of the 4th Corpus Linguistics Conference*, CL '07, Birmingham, UK.

Ted Pedersen, Siddharth Patwardhan, and Jason Michelizzi. 2004. WordNet::Similarity: measuring the relatedness of concepts. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (Demonstration Papers)*, HLT-NAACL '04, pages 38–41, Boston, MA.

James Pustejovsky, Peter Anick, and Sabine Bergler. 1993. Lexical semantic techniques for corpus analysis. *Computational Linguistics*, 19(2):331–358.

James Pustejovsky. 1995. *The Generative Lexicon*. MIT Press, Cambridge, MA.

Philip Resnik. 1993. *Selection and information: a class-based approach to lexical relationships*. Ph.D. thesis, University of Pennsylvania, UMI Order No. GAX94-13894.

Barbara Rosario and Marti Hearst. 2001. Classifying the semantic relations in noun compounds via a domain-specific lexical hierarchy. In Lillian Lee and Donna Harman, editors, *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*, EMNLP '01, pages 82–90, Ithaca, NY.

Barbara Rosario, Marti A. Hearst, and Charles Fillmore. 2002. The descent of hierarchy, and selection in relational semantics. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, ACL '02, pages 247–254, Philadelphia, PA.

Diarmuid Ó Séaghdha and Ann Copestake. 2009. Using lexical and relational similarity to classify semantic relations. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '09, pages 621–629, Athens, Greece.

Diarmuid Ó Séaghdha. 2008. *Learning compound noun semantics*. Ph.D. thesis, Computer Laboratory, University of Cambridge, Published as Computer Laboratory Technical Report 735.

Richard Socher, Brody Huval, Christopher D. Manning, and Andrew Y. Ng. 2012. Semantic compositionality through recursive matrix-vector spaces. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, pages 1201–1211, Jeju, Korea.

Karen Spärck Jones. 1983. Compound noun interpretation problems. In Frank Fallside and William A. Woods, editors, *Computer Speech Processing*, pages 363–381. Prentice-Hall, Englewood Cliffs, NJ.

Amanda Spink, Dietmar Wolfram, Major B. J. Jansen, and Tefko Saracevic. 2001. Searching the web: the public and their queries. *J. Am. Soc. Inf. Sci. Technol.*, 52(3):226–234.

Takaaki Tanaka and Timothy Baldwin. 2003. Noun-noun compound machine translation: a feasibility study on shallow processing. In *Proceedings of the ACL 2003 workshop on Multiword expressions*, MWE '03, pages 17–24, Sapporo, Japan.

Robert Trask. 1993. *A Dictionary of Grammatical Terms in Linguistics*. Routledge, NY.

Stephen Tratz and Eduard Hovy. 2010. A taxonomy, dataset, and classifier for automatic noun compound interpretation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 678–687, Uppsala, Sweden.

Peter Turney and Michael Littman. 2005. Corpus-based learning of analogies and semantic relations. *Machine Learning Journal*, 60(1-3):251–278.

Peter Turney. 2006a. Expressing implicit semantic relations without supervision. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, COLING-ACL '06, pages 313–320, Sydney, Australia.

Peter Turney. 2006b. Similarity of semantic relations. *Computational Linguistics*, 32(3):379–416.

Lucy Vanderwende. 1994. Algorithm for automatic interpretation of noun sequences. In *Proceedings of the 15th conference on Computational linguistics*, COLING '94, pages 782–788, Kyoto, Japan.

Beatrice Warren. 1978. Semantic patterns of noun-noun compounds. In *Gothenburg Studies in English 41, Goteburg, Acta Universtatis Gothoburgensis*.